The genetic and ecological landscape of plasmids in the human gut

Authors: Michael K. Yu[1,‡,*], Emily C. Fogarty[2,3,‡], A. Murat Eren[2,3,4,5,*]

Affiliations:

[1] Toyota Technological Institute at Chicago; Chicago, IL 60637, USA

[2] Department of Medicine, University of Chicago; Chicago, IL 60637, USA

[3] Graduate Program in the Biological Sciences, University of Chicago; Chicago, IL 60637, USA

[4] Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory; Woods Hole, MA 02543, USA

[5] Helmholtz Institute for Functional Marine Biodiversity, Ammerländer Heerstraße 231, 26129 Oldenburg, Germany

‡ These authors contributed equally to this work.

[*]Correspondence: mikeyu@ttic.edu, meren@uchicago.edu

Abstract

Plasmids are mobile genetic elements found across all domains of life. As plasmids often encode determinants of fitness, their evolution is intertwined with their hosts. However, naturally occurring plasmids remain far less understood than their hosts due to the lack of frameworks to recognize plasmids and to classify them into evolutionary groups. Here we trained a machine learning model that recognizes plasmids based on genetic architecture with state-of-the-art accuracy. We applied this model to a global collection of human gut metagenomes to identify 68,350 unique plasmids, a 200-fold increase over the number of known plasmids that could be detected in this environment. To understand the evolution of these plasmids, we developed a generalizable approach that enabled us to define 1,169 'plasmid systems'. Each system consists of plasmids that share a backbone sequence containing core plasmid functions, such as replication and conjugation, but vary in cargo genes that are often critical to the host, such as antibiotic resistance, amino acid biosynthesis, and tRNA modification. Members of the same system are often found in geographically distinct human populations, revealing cargo genes that likely respond to environmental selection. The ecological patterns of plasmids we observed could not be explained by microbial taxonomy. This work uncovers the tremendous diversity of plasmids and demonstrates the need to characterize them as a separate component of microbiomes distinct from their hosts.

Main

Plasmids are a type of mobile genetic element[1] that occur in all domains of life[2,3]. They typically exist as extrachromosomal and circular DNA, replicate semi-independently of their hosts, and often transfer between cells as a mechanism of horizontal gene transfer[4–8]. A hallmark of plasmids is their remarkably diverse capacity to impact their microbial hosts, by carrying fitness-determining functions[4–6] such as antibiotic resistance genes[9,10] and virulence factors[11,12]. Plasmids also exhibit many interesting genetic properties, such as frequent recombination, which can result in recurrent "backbone" sequences that are shared by multiple plasmids[13–15]. These backbone sequences often encode for core replication and transfer machinery[13,14,16–18] that can determine their host range[18,19] as well as copy number in a specific host[20]. Experiments in model systems and cultured organisms have revealed the critical impact of plasmids in microbial phenotypes and survival especially for pathogens with medical significance. Yet, our understanding of the diversity, ecology, and genetic architecture of naturally occurring plasmids are far from complete.

Recent advances in metagenomics offer unprecedented access to the entire DNA content of an environment without the need for cultivation[21]. In particular, metagenomic assembly and binning strategies have enabled the reconstruction and characterization of microbial genomes de novo[22], including those in the human gut[23] where microbes have been associated with health and disease states[24,25]. Metagenomic approaches have also been applied to study plasmid content[26], but this application has been limited to enriching for plasmid DNA through library preparation techniques or to surveying only a small handful of metagenomes at a time[27–32]. Over the past decade, the number of publicly available metagenomes has rapidly increased, numbering in the tens of thousands, creating an opportunity to study plasmids at an unprecedented scale in complex ecosystems.

Comprehensive insights into plasmid ecology and evolution require effective computational strategies for de novo identification of plasmids, which remains a challenge[33]. Several computational strategies have been developed to identify plasmids in sequence collections[17,34,35]. Many of these approaches rely on k-mer patterns learned from reference plasmid sequences[30,31,36], exploit known functions such as replication or conjugation genes[37–39], or use a combination of these features[35]. While these features can help identify plasmids similar to those in public databases, they are of limited utility to recognize novel plasmids. Other approaches focus on circularity of sequences during (meta)genomic assembly[32,40,41]; however, this strategy

overlooks plasmids that are linear, integrated, or found as assembly fragments, and may confuse other types of circular mobile elements for plasmids.

Here, we present (1) a machine learning approach to identify plasmids in complex microbial ecosystems, (2) and a novel algorithm to gain insights into plasmid evolution at scale. Specifically, we identify a collection of 68,350 non-redundant plasmids in the human gut microbiome that were more genetically diverse than reference plasmids and substantially more prevalent across global human populations. Using a novel network partitioning algorithm, we organize this large-scale sequence collection into 'plasmid systems' based on shared backbone sequences, and demonstrate that plasmid systems provide a framework for studying the selection of plasmids by environmental pressures.

A plasmid classification system based on *de novo* gene families

To enable a systematic study of plasmid sequences for machine learning, we compiled a reference set of 16,827 plasmids and 14,367 chromosomal sequences from public databases (Figure 1A, Table S1). In these sequences, we identified 51.2 million open reading frames and annotated them with functions defined in the Cluster of Orthologous Genes (COG)[42] and Pfam[43] databases. We also used MMseqs2[44] to organize genes *de novo* into 2,322,750 gene families and removed those that contained only one gene. The remaining 1,090,132 *de novo* families enabled a more comprehensive analysis by accounting for 95% of all plasmid genes (Figures 1B and S1A). Using *de novo* gene families also substantially increased our ability to identify genes that were enriched in plasmids, independently of available gene function databases (Figures S1B and S1C).

We used this reference database to train a machine learning model, PlasX, that distinguishes between plasmids and chromosomes based on genetic architecture (Figure 1C). PlasX is a logistic regression, which assigns a positive (or negative) coefficient to gene families that are likely to originate from a plasmid (or chromosomal) sequence. The coefficients of gene families within a sequence are summed to calculate a prediction score, ranging from 0 to 1, where a score of >0.5 designates a plasmid. To improve performance, PlasX uses a technique called elastic net regularization, which identifies gene families with redundant or noisy signals and then minimizes the usage of these families by setting their coefficients equal or close to zero. Consequently, only

a non-redundant and informative set of gene families can impact predictions by having coefficients far from zero (Figure S1D). For training and evaluating PlasX, we used 10kb slices of the reference sequences, to normalize for the fact that chromosomes are generally much longer than plasmids and to improve downstream application of PlasX on sequence collections that may contain a large number of fragmented sequences, such as assembled metagenomes.

Benchmarking the efficacy of a plasmid prediction algorithm is a non-trivial task. Evaluating an algorithm's performance on sequences that are similar to those used during training, or comparing it to older approaches that were trained on a much smaller number of sequences are common pitfalls that inflate accuracy estimates. Here we implemented a more realistic evaluation framework to compare PlasX to three state-of-the-art algorithms, PlasClass[31], PPR-Meta[45], and Platon[46]. We first evaluated performance in 4-fold cross-validation, using a 'naive' randomized splitting of sequences into training and test data (Figure S1E). PlasX achieved nearly perfect accuracy, with the highest area under the precision-recall curve (AUCPR=0.99) compared to all other methods (Figure S1F). While naive splitting is a common evaluation technique, it is not a fair strategy as it can separate very similar sequences into training and test data, especially given the redundancy of sequences in public databases, and thus inflate the accuracy of classification. As a more accurate benchmark, we (1) designed an 'informed' split by first clustering plasmid and chromosomal sequences into subtypes and then keeping all sequences in the same subtype together in either the training or test data to better evaluate the ability of recognizing novel sequences and (2) assigned normalized weights to sequences to prevent well-studied plasmids from influencing the prediction ability disproportionately (see Methods). This advanced benchmark revealed a greater performance divide between PlasX (weighted AUCPR=0.70) and all other methods, with the next best method performing substantially worse (Platon, weighted AUCPR=0.23) (Figure 1D). Finally, we evaluated the performance of PlasX and other methods on a novel and recently characterized plasmid of Wolbachia, pWCP[47]. Since pWCP was not present in the training data for any of the prediction methods, it provided a unique opportunity to investigate whether this plasmid, which remained elusive until recently, could have been discovered through a de novo plasmid survey. PlasX was able to predict pWCP as a plasmid (score = 0.73), while all other methods (PlasClass[31], PPR-Meta[45], Platon[46], and Deeplasmid[35]) were unable to classify it as a plasmid, either labeling it incorrectly as a chromosome or reporting high uncertainty in their prediction (Supplementary Table S8). Overall, these results suggest that PlasX, with its reliance on gene families rather than strictly defined sequence features, is unique in its ability to predict novel plasmids that are not present in existing databases.

PlasX's accuracy suggests it has learned insights into defining a "plasmid". A clear definition of a plasmid has remained elusive because commonly recognized features of plasmids can apply to other mobile or integrated genetic elements, including being extrachromosomal[32,48], facilitating horizontal gene transfer[49,50], or encoding traditional core functions like replication and mobilization[38,51]. To broaden our understanding of a plasmid, we used PlasX's coefficients to rank gene families by their importance in *de novo* identification of plasmids (Table S7). Among the 200 most important gene families, 19 were COGs and Pfams whose functional descriptions can be immediately recognized as being plasmid-associated because they contain keywords such as "plasmid", "replication", or "conjugation" (Figure 1E). However, another 9 COGs and Pfams did not have such a recognizable description. For example, a family of lipoproteins (PF05714) has been studied for conferring virulence in a few plasmids[52,53], but it is not generally thought of as a common plasmid function. Nonetheless, this family had the 17th highest coefficient of 1.678, consistent with its enrichment in 168 plasmids (36 plasmid subtypes) but only 2 chromosomes.

De novo families also provided two types of novel insights about plasmids. One insight is that 56.1% of *de novo* families can be thought as 'subfamilies' that group together a subset of genes within the same COG or Pfam. As many COGs and Pfams contain both plasmid and chromosomal genes, these subfamilies can provide a deeper resolution of the bacterial gene pool by delineating plasmid- or chromosome-evolved lineages. For example, the Pfam PF10609 is a broad family of genes related to *parA*, a gene that drives the partitioning of chromosomes[54] and plasmids[55] during cell division. As this family is found on 35% of plasmids and 95% of chromosomes, it alone is not informative for identifying plasmids and thus has a coefficient close to zero (-0.023). However, PF10609 can be further dissected into plasmid-specific subfamilies, such as mmseqs_5_1535552 (coefficient +0.455), and chromosome-specific subfamilies, such as mmseqs_70_40217271 (coefficient -0.198). Indeed, the maximum likelihood phylogenetic tree that relates the genes in these two subfamilies show a divergence of plasmids and chromosomes into monophyletic groups (Figure 1F), which is also reflected in their sequence alignment (Figure 1G). The second insight is that 35.5% (398,174) of *de novo* families have no overlap with any COG or Pfam. Many of these families have highly positive coefficients (e.g. 12,076 families have coefficients >0.1) that make a sequence appear substantially more like a plasmid and thus could represent fundamental but unexplored plasmid functions.

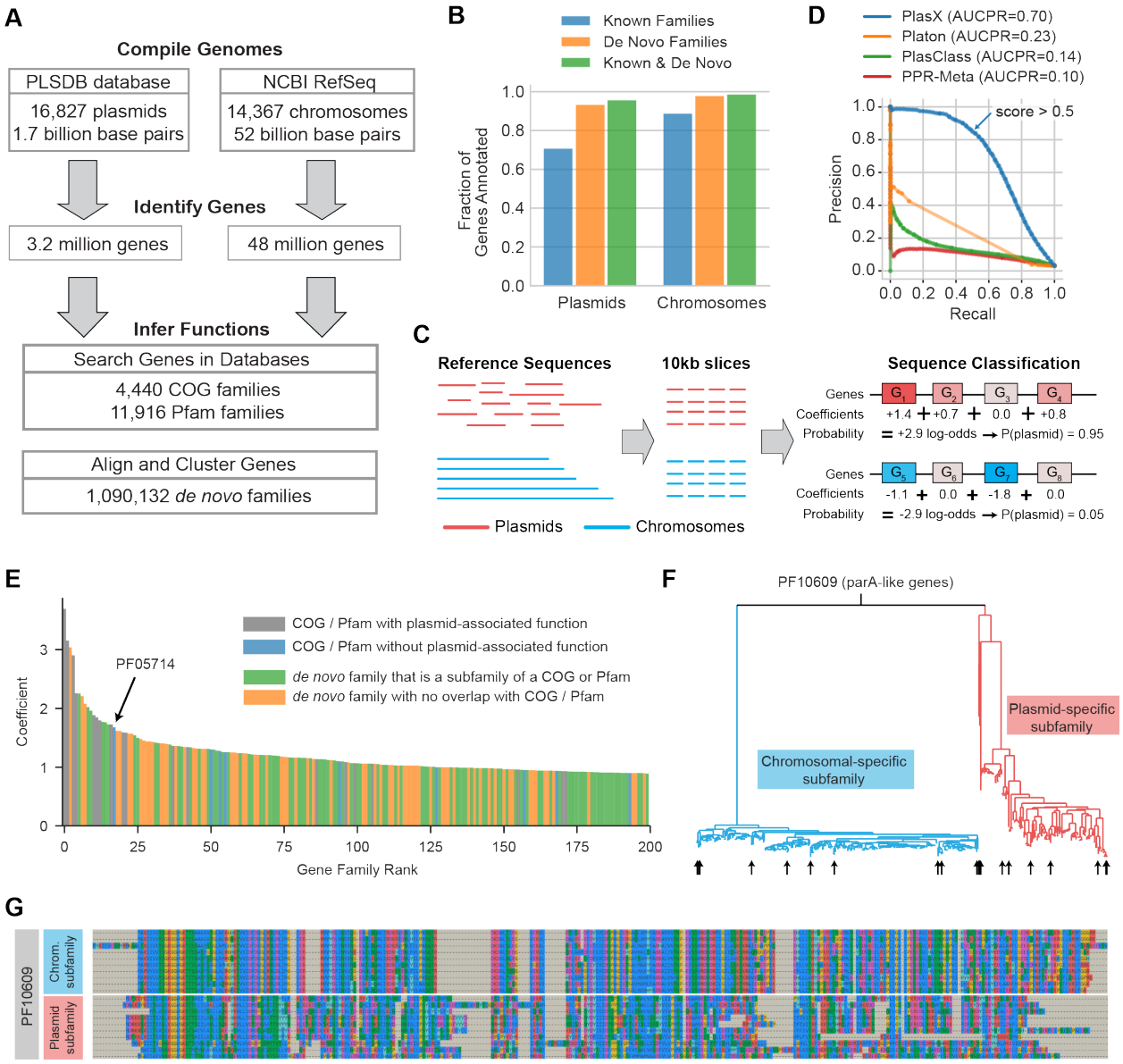


Figure 1. A machine learning model for classifying plasmids. (A) Our pangenomics workflow to characterize gene functions in a reference set of plasmids and chromosomes. **(B)** The fraction of all plasmids or all chromosomal genes that are annotated by using known families (blue), *de novo* families (orange), or a combination of both (green). **(C)** Training of PlasX. Reference sequences are sliced into 10kb windows and then prediction scores are made by a logistic regression that sums the contributions of gene families within a sequence. **(D)** Precision-recall curves comparing PlasX, Platon, PlasClass, and PPR-Meta. Except for PPR-Meta, every method was trained and evaluated using 4-fold cross-validation and an informed split. AUCPR was calculated using sequence weights for normalization. The arrow indicates the performance of PlasX using a score threshold of >0.5. **(E)** The 200 gene families with the highest PlasX coefficients and thus most important for identifying plasmids. Gene families are ranked by their coefficient. **(F)** Maximum-likelihood phylogenetic tree of genes that are in PF10609 and also in either the plasmid-specific *de novo* subfamily mmseqs_5_1535552 (red) or chromosome-specific *de novo* subfamily mmseqs_70_40217271 (blue). **(G)** Sequence alignment of 10 representative genes from each subfamily (arrows in F).

PlasX unveils a large database of new plasmids from the human gut microbiome

Having verified PlasX's ability to identify plasmid sequences, we applied it to survey naturally
170 occurring plasmids in the human gut microbiome, an environment which harbors a diverse range of microbes and mobile genetic elements[56]. We assembled 36 million contigs from 1,782 human gut metagenomes, spanning culturally and geographically distinct human populations (Table S2). Running PlasX on these data resulted in a total of 226,194 predicted plasmids with a model score above 0.5 (Figures 2A and S2A, Table S3). Our predictions spanned a wide range of lengths,
175 including 135 sequences that were longer than 100 kbp, but they were generally shorter than reference plasmids with a median length of 2.6 kbp versus 53.3 kbp, respectively (Figure S2B). This discrepancy can be partly explained by fragmentation during metagenomic assembly, as the median length of the entire set of contigs was 2.1 kbp, and only 50,310 (0.14%) contigs were longer than 100 kbp. To minimize this issue, we removed predictions that were likely assembly
180 fragments because they were subsequences of other predictions in our collection and also did not appear to be circular elements. This filter retained 100,719 predictions for downstream analyses.

To determine the circularity of predicted plasmids, we analyzed the orientation of paired-end
185 reads recruited from metagenomes. This is a powerful strategy because if a contig occurred as a circular element in the environment, then matching reads from the same pair would be recruited to opposite ends of the contig in a 'reverse-forward' orientation, instead of the typical 'forward-reverse' orientation (Figure 2B). With this approach we found that 19,652 plasmid sequences were circular, and we designated them as high-confidence plasmids for downstream analyses.
190 These circular plasmids spanned a range of sizes, with a median length of 4.4 kbp and 854/378/47 plasmids longer than 25/50/100 kbp. An additional 14,151 sequences were not circular themselves but were highly similar to a circular sequence. Together, these two types of sequences defined a set of 'circular-associated' sequences representing 33.6% (33,803/100,719) of predictions. Multiple factors can explain the lack of signal for circularity for the remaining
195 plasmids, including insufficient sequencing depth to observe reverse-forward pairs, fragmented contigs, or the non-circular nature of some plasmids that occur linearly[57] or are integrated in a chromosome[3].

Beyond circularity, confirming *in silico* whether a novel sequence represents a plasmid is a
200 significant challenge. In the absence of single-copy core genes that have been vital to assess the completeness of non-plasmid and non-viral genomes assembled from metagenomes[22], our understanding of the canonical features of plasmids is limited to a relatively small set of well-studied genes that are primarily derived from plasmids of model organisms in culture[37,38]. For instance, MOB-suite[38] identified canonical features for plasmid replication and conjugation in only
205 16.3% of the 16,827 reference plasmid sequences used to train PlasX. This relatively small percentage reveals the limits of conventional approaches to identify plasmid features and thus foreshadows their limited utility to survey novel plasmids. Indeed, MOB-suite identified canonical features in only 10.1% of our predictions. Given this narrow sensitivity, we developed orthogonal data-driven strategies to measure confidence in our predictions.

210

For the remaining 89.8% (90,446/100,719) of predicted plasmid sequences in which MOB-suite did not find any canonical plasmid features, we performed several types of analyses to assess how many are true plasmids or novel sequences (Figure 2C). We found that 24.5% (24,689) of predictions were circular-associated sequences. In addition, 26.7% (26,921) were 'keyword-
215 recognizable', as they contained a COG or Pfam function with the words 'plasmid' or 'conjugation'. And finally, 4.0% (3,996) were highly similar to a known plasmid sequence in NCBI, while 64.7% (65,117) were novel sequences with no hits to any sequence in NCBI. Combining these individual statistics, we found that 49.4% (49,739) of predictions had some orthogonal support for being a plasmid, by MOB-suite or any of the first three types of analyses, and 28.5% (28,658) had such
220 support and were novel. Overall, these findings suggest that our collection of predicted plasmids include not only sequences that match known plasmids, facilitating the study of their diversity and gene pool in natural habitats, but also novel sequences that further advance our knowledge of plasmid biology.

225 We further investigated the subset of predictions that were highly similar to a sequence in NCBI and categorized matches as either known plasmids (26.9%), chromosomes (21.3%), viruses (0.6%), or an unclear type of sequence (51.2%). A total of 189 predictions matched a known virus. Of these, 110 were recognized as plasmids by MOB-suite or keywords but also contained virus-related COG or Pfam functions, as indicated by the keywords 'virus', 'viral', and 'phage'. These
230 predictions carry both plasmid and viral features, a phenomenon that has previously been reported[58–61]. Surprisingly, 808 predictions that matched a known chromosome were also circular-associated and recognized by MOB-suite or plasmid keywords. One explanation of these data is

that these plasmids can switch between an extrachromosomal or a chromosome-integrated state. This dual state is also present in integrative and conjugative elements (ICE), another type of mobile element that are typically replicated while integrated into the chromosome[62]. These plasmids that share similarity to viruses or ICEs are consistent with previous perspectives that the boundary between traditional categories of mobile elements can be blurry[8,48,63,64].

In general, it is likely that some of our predictions are not plasmids. To address such model errors, we observed that predictions matching known chromosomes in NCBI had lower model scores (<0.9), while those matching a reference plasmid were enriched in higher model scores (>0.9) (Figure S2C). This result demonstrates that PlasX's model score is well calibrated such that a stricter threshold, such as >0.9 instead of our current choice of >0.5, could be applied to filter for more confident predictions at the cost of sensitivity (Figure S2A).

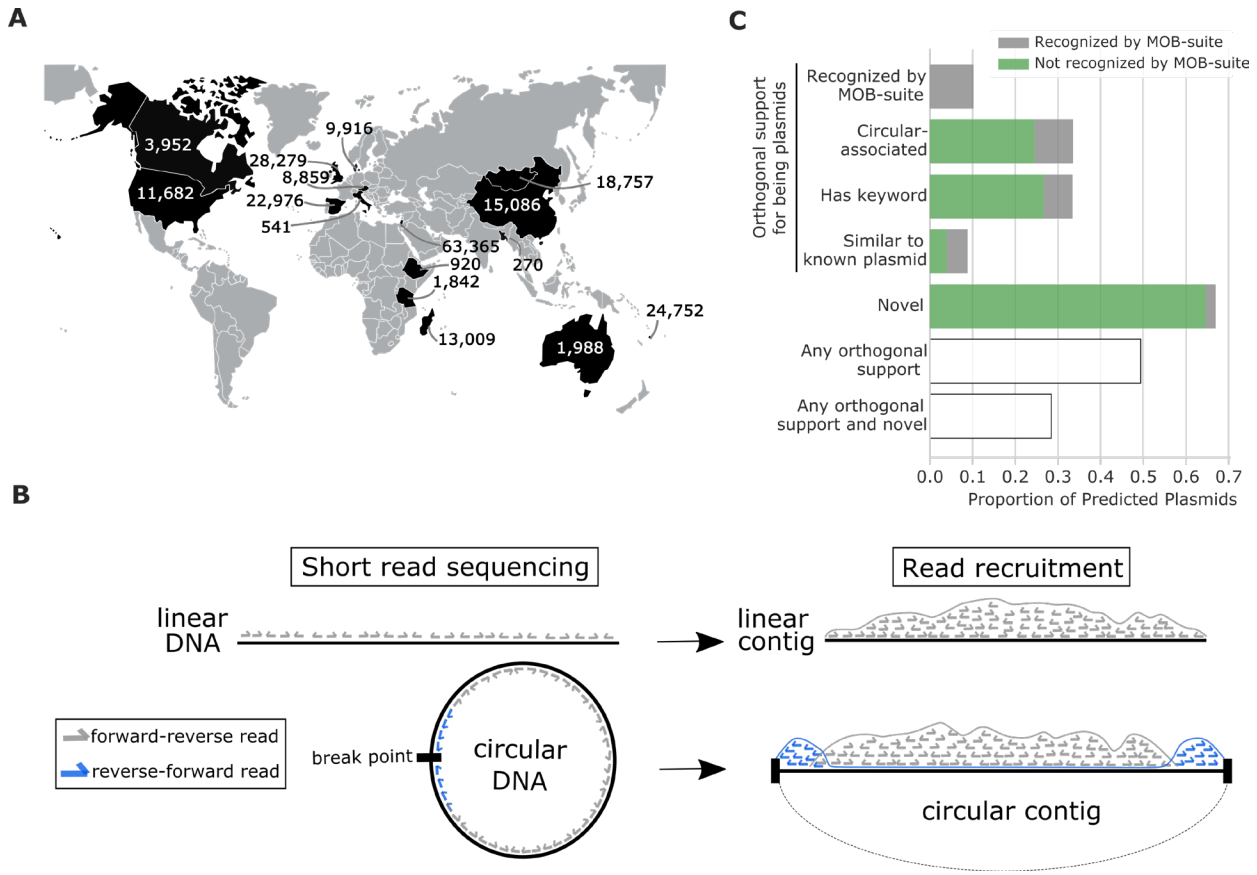


Figure 2. Plasmid prediction from metagenomes. (A) Number of plasmids predicted from each country. **(B)** Diagram of paired-end reads mapping to a linear versus a circular contig. Linear contigs have forward-reverse reads only, while circular contigs also have reverse-forward reads concentrated on the ends due to an artifact in contig assembly. **(C)** Orthogonal support for and novelty of the 100,719 non-fragment predictions.

Plasmids predicted from metagenomes are found in isolate genomes and can transfer between microbial populations

To experimentally validate our metagenome-derived predictions as true plasmids of the human
255 gut, we developed a pipeline for identifying predictions that are (1) present in human gut microbial isolates, (2) are circular in those isolates, and (3) can be naturally transferred to other microbes. First, we detected 127 of our predicted plasmids in 14 *Bacteroides* isolate genomes that we sequenced in a previous study[65] (Figure 3A). Short-read sequencing of two of these isolates suggested that the predicted plasmids pFIJ0137_1 and pENG0187_1 were circular based on
260 paired-end orientation (Figure 2C). We further confirmed their circularity using long-read sequencing. Following a previously described approach[47], we identified and manually confirmed 500 long reads that align completely to a plasmid but not to the host chromosome (Figures 3A and S3). Some of these long reads align across the artificial contig breakpoint, indicating these plasmids are extrachromosomal and circular (see Methods).

265

Finally, we tested the ability of pFIJ0137_1 to transfer between its host, *B. fragilis 214* (one of 14 isolates from [65]) to a well-known laboratory strain, *B. fragilis 638R.* We designed an experimental setup that takes advantage of the naturally encoded erythromycin resistance (erm^R) on pFIJ0137_1 and the rifampicin resistance (rif^R) of *B. fragilis 638R*. Specifically, we first mated
270 isolates in the absence of antibiotics, and then selected for transconjugants on media containing both antibiotics (Figure 3B). While this plasmid lacks conjugation machinery, it contains two relaxases (blue genes in Figure 3A) and thus could be mobilized by different conjugative apparatus in the host cell. Through short-read sequencing of the donor, recipient, and resulting transconjugants, and by employing a read recruitment analysis, we confirmed that pFIJ0137_1
275 transferred from *B. fragilis 214* to *B. fragilis 638R* (Figure 3C). This analysis also confirmed the circularity of pFIJ0137_1 in *B. fragilis 214* and both *B. fragilis 638R* transconjugants. Besides a 68bp deletion, pFIJ0137_1 in *B. fragilis* 214 (isolated in Chicago, USA) was identical to the pFIJ0137_1 version assembled from a Fijian metagenome, suggesting a relatively recent transfer of this plasmid between unrelated human populations. These experimental results show that while
280 PlasX identifies plasmids solely based on genetic architecture, it is capable of predicting plasmids that have canonical features of being extrachromosomal, circular, or transmissible between cells.

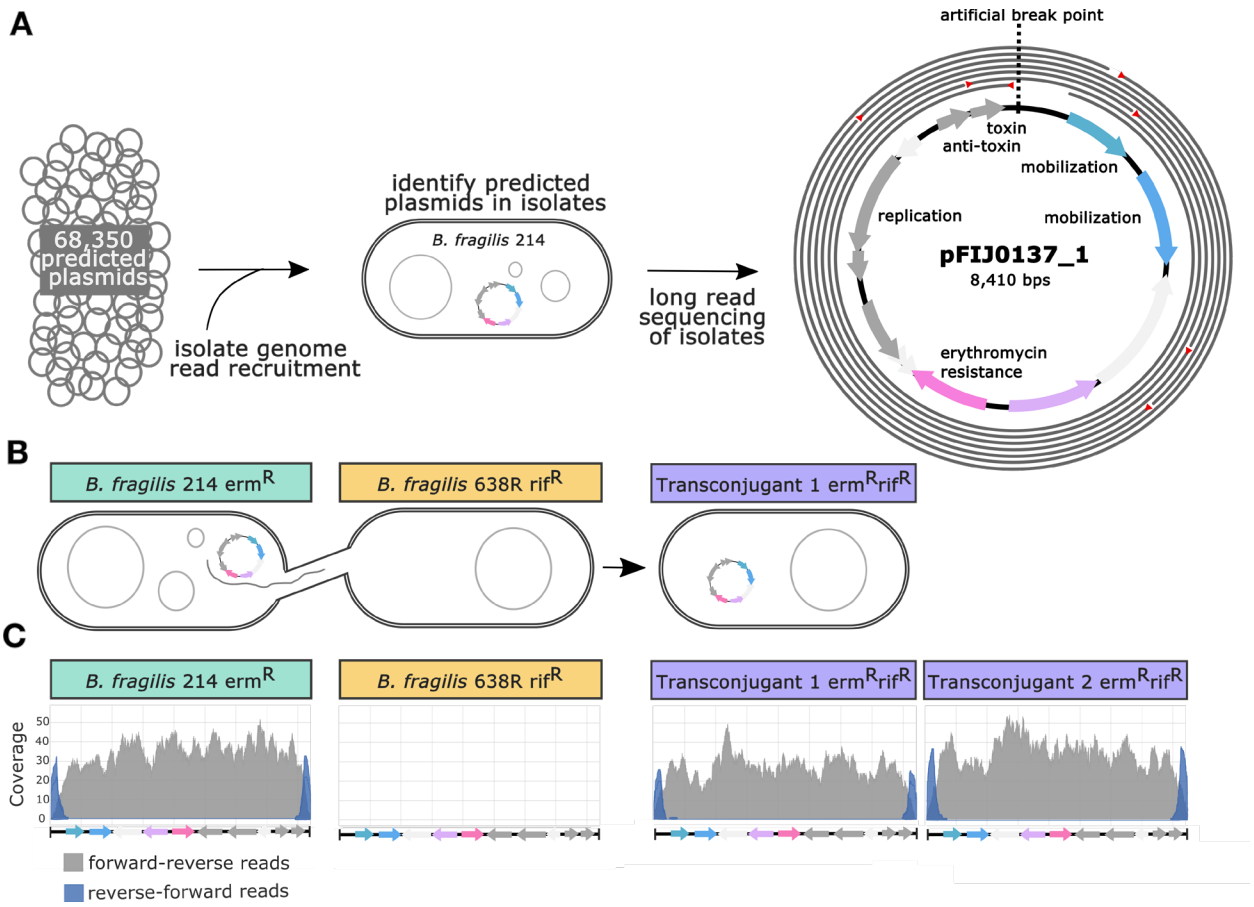


Figure 3. Experimental validation of plasmid predictions. (A) We recruited reads from the sequenced genomes of 14 *Bacteroides* isolates to determine which isolates contain our predicted plasmids. We further confirmed the presence and circularity of a predicted plasmid (pFIJ1037_1) in the isolate *B. fragilis* 214 by long read sequencing. Grey circles represent 7 (of 500) long reads that align to pFIJ0137_1. Red triangles designate the beginning of a long read. **(B)** Transfer of pFIJ1037_1 from *B. fragilis* 214 to *B. fragilis* 638R via conjugation and selection on erythromycin- and rifampicin-containing media. **(C)** Coverage plots showing read recruitment of *B. fragilis* whole-genome sequencing reads to the pFIJ1037_1 reference sequence, confirming transfer of pFIJ1037_1. Grey are forward-reverse reads, while blue are reverse-forward reads that indicate the circularity of pFIJ1037_1.

Novel plasmids are highly prevalent, reflect human biogeography, and unexplained by microbial taxonomy

Next, we sought to characterize the ecology of plasmids across human populations through
300 metagenomic read recruitment. For this task, we first dereplicated the entire collection of reference and predicted plasmid sequences, where we assumed that any pair of plasmid sequences was redundant if at least 90% of either sequence aligned to the other with over 90% sequence identity. This analysis found 68,350 and 11,121 non-redundant sequences in the set of 226,194 predicted and 16,827 reference plasmids, respectively. Then, we used the non-
305 redundant sets of plasmids to recruit reads from the 1,782 globally distributed human gut metagenomes. We labeled a plasmid as present in a metagenome if its 'detection' was ≥0.95, where detection is the fraction of the sequence covered by at least one read (see Methods).

Our read recruitment analysis revealed that predicted plasmids were much more prevalent across
310 human populations than reference plasmids. For instance, only 1.9% (211) of reference plasmids were present in at least two individuals in our dataset, suggesting the limited ecological relevance of reference plasmids to naturally occurring gut microbial communities. Indeed, many reference plasmids were isolated from a relatively small number of pathogens, such as *Escherichia coli*, *Salmonella enterica*, *Pseudomonas aeruginosa*, *Klebsiella pneumoniae*, and *Vibrio cholerae*,
315 which are unlikely to be abundant in healthy humans. In contrast, 63.1% (43,114) of the predicted plasmids were present in at least two individuals (Figure S2D). Moreover, of the most highly prevalent plasmids found in ≥100 individuals, 99.7% (5,400/5,414) were predicted plasmids while only 0.3% (14/5,414) were reference plasmids.

320 The prevalence of predicted plasmids suggests that they capture the biogeography and lifestyles of human populations more effectively than reference plasmids. To confirm this, we performed agglomerative clustering to construct a dendrogram that organizes metagenomes based on their plasmid content. Using reference plasmids for this clustering, we found that only 50.2% of metagenomes were arranged next to another metagenome from the same country. In contrast,
325 using predicted plasmids (Figure 4B) resulted in 74.0% of metagenomes arranged that way. We also organized metagenomes using a dimensionality reduction of predicted plasmids. This analysis shows that industrialized versus non-industrialized metagenomes can be distinguished solely by their plasmid content (Figure 4C). Dimensionality reduction also showed country-specific clustering (Figure S4).

330 These results have parallels with previous studies that found associations between gut microbiota, as characterized by microbial taxonomy, and the geography and lifestyles of human populations[66–69]. As geography is correlated to both plasmids and taxonomy, we wondered how many of our 68,350 plasmids are ecologically associated with and therefore explained by taxonomy. On one hand, such associations may be strong because plasmids are symbionts that
335 rely on host machinery for replication and can have a narrow host range. On the other hand, such associations may be weak or nonexistent for two reasons. Some plasmids are known to have a range of multiple hosts[70–72], which might not be neatly defined by a single species or even higher taxonomic category such as genus or phylum. Additionally, plasmids are often nonessential elements that can be gained or lost, such that nearly identical microbes can differ by the presence
340 or absence of a plasmid or in the number of plasmid copies. Here, we systematically examined the ecological associations between plasmids and taxonomy to determine if plasmids comprise an independent component of microbial systems.

For every plasmid, we inferred its most likely host as the taxonomic group that had the most
345 similar ecological distribution (see Methods). We surveyed taxonomic groups across all levels, from subspecies and species to class and phyla. We used two different formulas to calculate the ecological similarity between a plasmid and potential host: (1) the correlation in the abundance levels of the plasmid and host across metagenomes, and (2) how often the plasmid and host are found together in the same metagenome. Although some predicted plasmids had a high
350 ecological similarity with their best matching taxonomic group, the vast majority of predicted plasmids had low similarity scores (median correlation = 0.04, median Jaccard = 0.21) (Figures S5A and S5B). We also observed low similarity scores even for reference plasmids that are isolated from a defined microbial host (Figures S5C and S5D). For example, the plasmid pDOJH10S and its cognate host, *Bifidobacterium longum*, were present together in 10
355 metagenomes; however, 27 and 69 metagenomes contained only the plasmid or only the host, respectively (Figure S5E).

Overall, our findings suggest that plasmids are a highly complex and prevalent feature of microbiomes (Figures 4A, 4B, and S2D), forming an ecological dimension that can stratify human
360 populations (Figures 4C and S4) but cannot be explained by microbial taxonomy alone (Figure S5). While high-throughput analyses of human gut microbiomes often focus on taxonomic features, it has been challenging to find significant or reproducible taxonomic associations that distinguish health and disease states[73,74]. As plasmids often carry key determinants for survival

in an environment, we propose that systematic analysis of plasmid ecology is necessary to develop a complete understanding of the human microbiome.

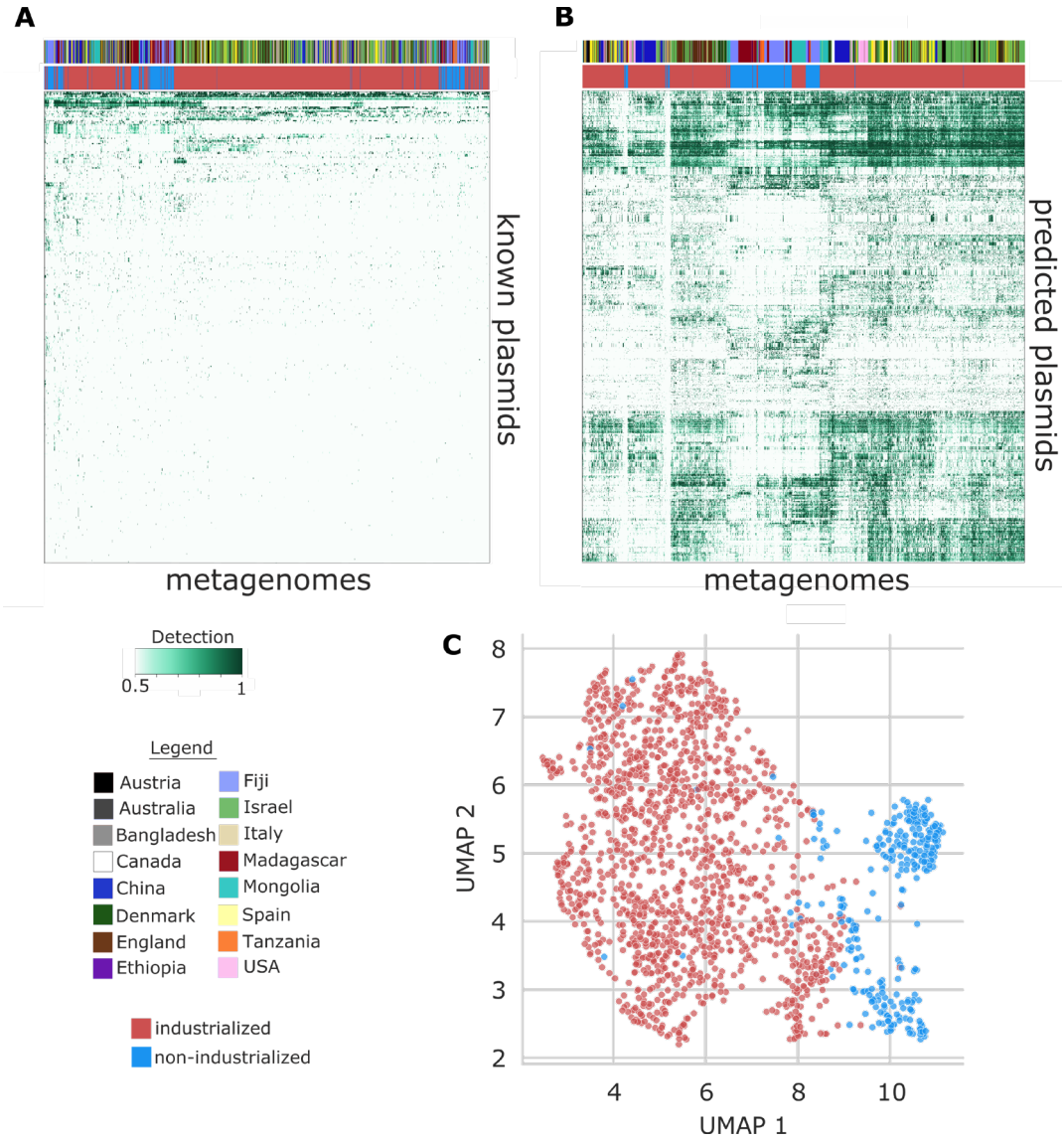


Figure 4. Global plasmid ecology. (A) Read recruitment of human gut metagenomes to 11,121 non-redundant reference plasmids. The heatmap shows the 338 plasmids that are present in at least one metagenome (≥0.95 detection). **(B)** Read recruitment to 68,350 non-redundant predicted plasmids. The heatmap shows the 1,000 most prevalent plasmids that are present in at least one metagenome and have PlasX score ≥0.75. In A and B, column colors indicate country of origin and lifestyle (industrialized or non-industrialized). **(C)** Clustering of metagenomes based on the predicted plasmids that are present, using the UMAP dimensionality reduction method[75]. Metagenomes from industrialized or non-industrialized populations are colored red or blue, respectively.

Plasmid systems organize evolutionarily related plasmids by distinguishing backbone versus cargo content

Our large collection of predicted plasmids provides an unprecedented opportunity to study evolutionary patterns in plasmids and the extent to which they occur ecologically. Due to frequent genetic rearrangements, a hallmark of plasmid evolution is the reuse of a backbone and emergence of varying cargo/accessory genes[13,15,17,18]. The backbone typically encodes machinery necessary for plasmid maintenance, while the cargo represents additional genetic content, such as antibiotic resistance or other fitness-determining functions. While backbones can be examined experimentally, most studies have identified them computationally. Nonetheless, there are four major challenges to this computational task. First, there is a lack of consensus across studies on how to identify a plasmid backbone, with varying definitions based on nucleotide identity[13,14,77], gene similarity[18], or gene annotations[78–80]. Second, these methods do not verify that an identified backbone encodes a sufficient set of functions for plasmid replication. Third, these methods are typically designed to analyze a small set of plasmids in a single study or dataset. Finally, scaling methods to identify backbones in metagenomic data introduces extra complications related to plasmid redundancy and assembly fragments that could inflate the number of predicted backbones.

We designed a scalable algorithm called MobMess (Mobile Element Systems) to study backbone structure in our collection of plasmids. Compared to previous methods, MobMess has the advantage of being able to simultaneously compare sequences without relying on gene annotations and to handle metagenomic issues of redundancy and fragmentation (see Methods). First, MobMess calculates pairwise alignments across all plasmids to build a sequence similarity network, in which a directed edge represents the containment of one plasmid within another (defined by ≥90% sequence identity and ≥90% coverage of the smaller plasmid) (Figures 5A and S8). Next, MobMess recognizes and collapses redundancy between plasmids. Finally, MobMess analyzes patterns of connectivity in the network to define and identify ‘backbone plasmids’ that satisfy two criteria. First, the backbone plasmid must be a circular element, inferred here by paired end orientation (Figure 2B), to ensure that it is not an assembly fragment and, importantly, that the genes present are sufficient for plasmid replication. Second, a backbone plasmid must be found as a subsequence within one or more ‘compound plasmids’. These compound plasmids are composed of the backbone and additional cargo, indicating the ability to acquire or lose genes.

Together, we define a backbone and its compound plasmids as an evolutionary unit called a
410 'plasmid system' (Figure 5A).

This definition of plasmid systems enables a formal categorization of plasmids into evolutionarily cohesive groups and facilitates analyses of backbone versus cargo content and their ecology, much in the same way that pangenomes enable studies of core versus accessory gene content
415 in microbial genomes. However, plasmid systems are a specific case of pangenomics, as it is unlikely to find a naturally occurring microbial genome composed only of core genes. In contrast, backbone plasmids represent a minimal entity that can propagate using only backbone genes. MobMess provides an automated framework and standardized vocabulary to study this concept. In addition, MobMess makes accessible a set of tunable parameters, to methodically explore the
420 concept of plasmid systems at scale across different studies and datasets.

Other methods have recently been developed to cluster thousands of plasmids[81,82], but unlike MobMess, they are not designed to identify plasmid systems or analyze metagenomic data. To compare methods, we ran MobMess on the same set of 9,894 reference plasmids analyzed by
425 Redondo-Salvo et al[81] (Figure S6). In their study, Redondo-Salvo et al. constructed a plasmid similarity network with 79,727 edges. However, these edges span a wide range of similarity levels, where 66.5% of edges represent an alignment that covers <90% of either sequence (≥10% is not aligned) and 19.0% of edges have <70% alignment coverage (≥30% is not aligned). In contrast, MobMess applies a stricter threshold of ≥90% coverage to construct a smaller but more refined
430 set of 39,680 edges (connecting 25,270 unique pairs of plasmids). Moreover, Redondo-Salvo et al.'s edges are undirected, while MobMess's edges are directed to track smaller versus larger sequences. Retaining this extra information allows MobMess to distinguish between the 10,860 pairs (43.0%) with unidirectional connections, representing a backbone contained in a compound plasmid, versus the 14,410 pairs (57.0%) with bidirectional connections, representing nearly
435 identical plasmids.

Besides network construction, these methods also diverge in how they conceptually organize plasmids. MobMess dereplicates the 9,894 plasmids into 7,132 non-redundant sequences and then organizes them into 1,044 plasmid systems. In contrast, Redondo-Salvo et al. identified 641
440 clusters, which they call 'PTUs'. We found that 135 PTUs did correspond one-to-one to a plasmid system in MobMess, but the other PTUs spanned a wide range of evolutionary relations. At one extreme, 251 PTUs were simple sets of nearly identical plasmids, representing recent and strong

relations. At the other extreme, 45 PTUs were complex mixtures of distinct plasmid systems, representing distant and weak relations. For example, the largest PTU contained 2,460 plasmids, which MobMess further dissected into 1,481 non-redundant plasmids and 461 plasmid systems. Figure S7 visualizes one of these plasmid systems—MobMess precisely connects the system's backbone to its compound plasmids in a "star"-like topology, while Redondo-Salvo et al. connects almost every pair of these plasmids in a "hairball" that obfuscates the system's internal organization. Beyond this dataset, Redondo-Salvo et al. and another related method by Acman et al.[82] have only been tested on reference plasmids that have been completely assembled, while the unique algorithmic design of MobMess is designed to handle metagenomic data by distinguishing between fragmented versus complete (circular) plasmids.

MobMess identifies 1,169 plasmid systems with conserved backbones and a wide repertoire of cargo functions

We ran MobMess on our predicted plasmids and identified a total of 1,169 plasmid systems, naming them PS1 (plasmid system #1) to PS1169. While plasmid systems captured a small fraction of the genetic diversity among non-redundant plasmids (6.5%, or 4,424/68,350), they captured a large fraction of all circular plasmid contigs (72.7%, or 14,285/19,652) (see Methods, Table S4). The remaining 26.9% (5,367/19,652) of circular plasmids might be part of plasmid systems in nature, but MobMess did not categorize them as so because a corresponding backbone or compound plasmid was not present in our data.

Plasmid systems were highly heterogeneous in their genetic complexity. 602 plasmid systems contained at least 2 non-redundant compound plasmids, with the largest system containing 168 non-redundant compound plasmids (Figure 5C). For example, pFIJ1037_1, the plasmid we isolated and transferred between *B. fragilis* organisms, was part of PS486, a system containing 24 non-redundant plasmids and found across a total of 127 metagenomes. PS486's backbone consists of a replication protein and a toxin-antitoxin system, and the cargo genes include beta-lactamases, erythromycin resistance, tetracycline resistance and riboflavin biosynthesis (Figure 5D, Table S5).

PlasX often assigned higher model coefficients to backbone genes in the non-redundant set of predicted plasmids, suggesting these genes define the 'essence' of a plasmid by encoding essential functions that promote the ability of a plasmid to exist as a distinct element from the chromosome, such as the genes for plasmid replication, *repA* (PF01051), and mobilization, *mobA* (PF03432) (Figure 5B). In contrast, PlasX assigned lower coefficients to cargo genes, suggesting they encode functions that are not universally essential but important for specific niches, such as nitrogen reductase, *nifH* (PF00142), and membrane transport, *ompA* (PF00691). Indeed, 24.1% (2,169/8,995) of backbone genes versus 13.1% (4,514/34,526) of cargo genes encoded COG and Pfam functions with descriptions related to plasmid replication, transfer, and maintenance (see Methods).

The most frequent type of function encoded on cargo genes was antibiotic resistance, including efflux pumps, which can provide general resistance to multiple antibiotics, and genes targeting specific classes of antibiotics, such as glycopeptides and beta-lactams (Figure 6A). This large-

scale observation is consistent with numerous examples of known plasmids encoding resistance and further illustrates how the widespread presence of these plasmids pose a public health threat[85–88].

490

Other highly prevalent cargo functions included a wide diversity of cellular and metabolic pathways defined in the COG (Figure 6B) and KEGG databases (Figure S9). The most enriched among these was tRNA modification, encoded in 35 compound plasmids within different systems. For example, the globally prevalent system PS1110 (present in 739 metagenomes) contained
495 291 compound plasmids (27 non-redundant), three of which encoded an enzyme that performs tRNA Gm18 2'-O-methylation (COG0566) and were collectively present in 498 metagenomes (Figure S10, Table S5). This enzyme is thought to reduce the immuno-stimulatory nature of bacterial tRNA, which is detected by Toll-like receptors (TLR7) of the mammalian innate immune system[89,90]. While plasmids in some pathogens are known to facilitate bacterial evasion of
500 mammalian immune system by regulating surface proteins[91], the overwhelming prevalence of tRNA modification enzymes in our data suggests the likely presence of a previously unappreciated role for plasmids to increase the fitness of their bacterial hosts against the surveillance of the human immune system. Distinguishing between the fundamental structure of a plasmid (backbone genes) versus the genetic currency that is exchanged (cargo genes) allowed
505 us to organize the extensive plasmid diversity within plasmid systems and to recognize the recurrent evolution of the same cargo functions across different systems.

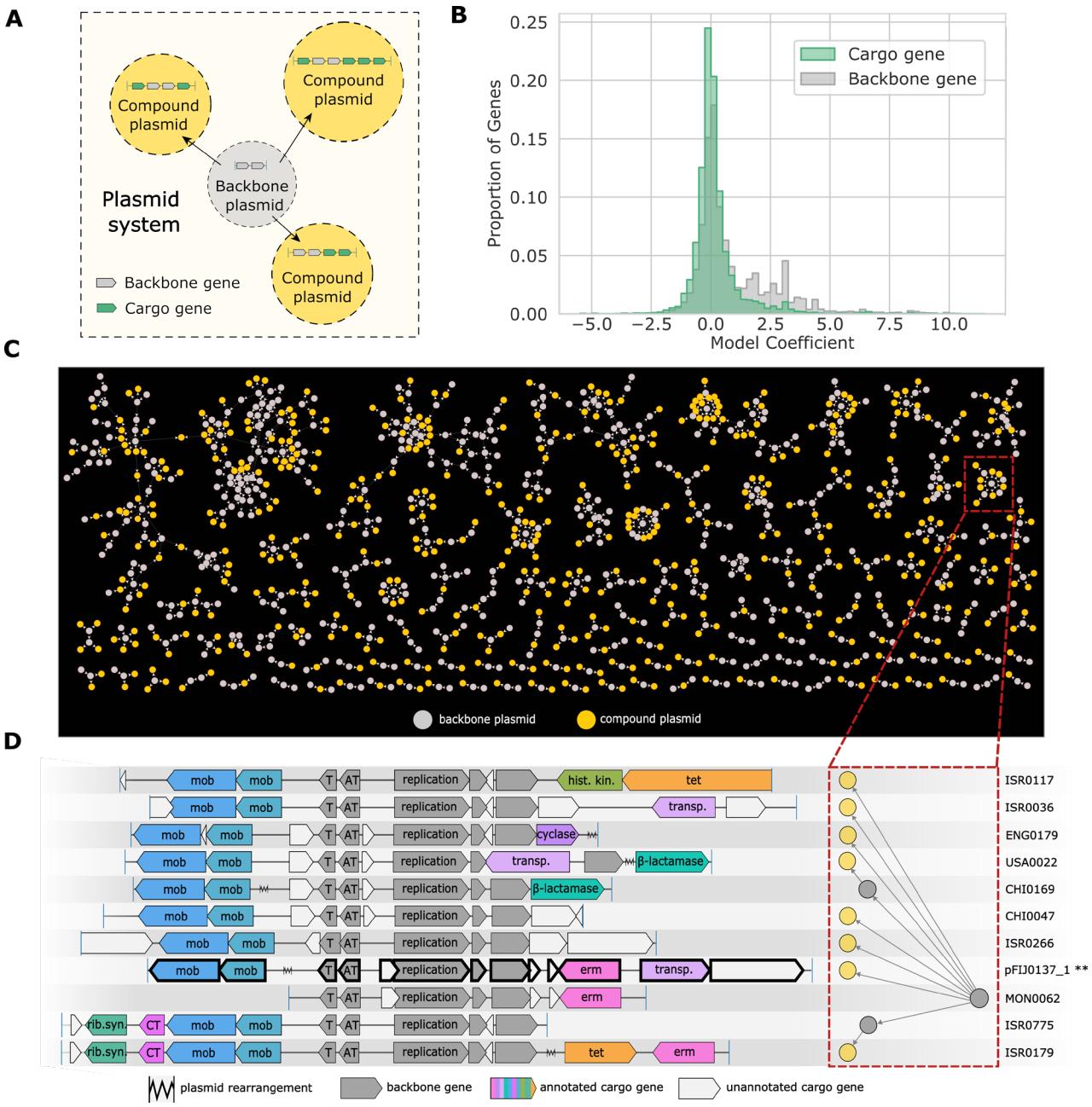


Figure 5. **Identification of plasmid systems. (A)** Network diagram of a plasmid system. **(B)** Distribution of model coefficients for backbone vs. cargo genes in the non-redundant set of 68,350 predicted plasmids. We excluded genes that lacked gene family annotations and thus have a coefficient of zero by default. We also excluded genes that were labeled as backbone with respect to some systems but cargo in others. **(C)** Network of all plasmid systems that contain ≥3 non-redundant and high-confidence plasmids. Only these types of plasmids are shown. **(D)** Genetic architecture of plasmids in PS486, encased by a red box in C. Two plasmids in C are excluded. The system's backbone (assembled from metagenome MON0062) encodes 5 backbone genes (colored gray). Rib.syn.=riboflavin biosynthesis, CT=conjugative transfer, mob=mobilization, T=toxin, AT=anti-toxin, tet=tetracycline resistance, erm=erythromycin resistance, transp.=transposon, hist. kin.=histidine kinase.

Plasmid systems adapt their cargo genes to specific environments

Thus far, we have observed that our collection of plasmids is highly heterogeneous in their ecological distributions (Figure 4B), yet they can also be organized by evolutionary relations into plasmid systems. To understand how ecology and evolution are intertwined phenomena, we asked whether plasmid systems span a single ecological niche or multiple niches. We assumed that every country represents a different niche, as countries are known to differ in microbial composition[69,92–97] and we observed that countries also differ in plasmid composition (Figure S4). We found that while individual plasmids are often present in a single country, a plasmid system frequently spans multiple countries (Figure 6C, Table S4). Indeed, 2,005 individual plasmids within a system were unique to a single country, yet 1,794 (89.5%) were part of more geographically diverse systems that were present in at least two countries. In fact, 84.0% (982/1,169) of systems were present in at least two countries, and 9 systems were even in 15 of the 16 countries represented in our data. We also found that plasmid systems are typically mutually exclusive, i.e. plasmids from the same system generally do not occur in the same individual human hosts. Consequently, metagenomes often have at most one plasmid from every system (see Methods, Figure S11A).

One explanation of this mutual exclusivity pattern is that a person may have never been exposed to multiple plasmids from the same system. This scenario is plausible for rare systems (e.g. present in <10 metagenomes). Indeed, we observed a trend where more prevalent systems exhibit less mutual exclusivity (R^2=0.39) (Figure S11B). However, we identified many prevalent systems (e.g. present in >100 metagenomes) that are more mutually exclusive than expected. For instance, PS961 is almost perfectly mutually exclusive across 131 metagenomes (Figure S11C). For such prevalent systems, a more likely explanation lies in either the backbone or cargo content of a system. The backbone typically encodes genes for replication machinery, and competition for these resources within a cell can lead to incompatibility between plasmids of the same system[98]. With a large number of plasmid sequences at hand, our dataset is a new resource that can support experimental investigations of the impact of incompatibility on plasmid replication[99,100], especially its contribution relative to external selective pressure on cargo genes. However, here we focused our attention on investigating ecological associations between environmental selection and cargo gene content.

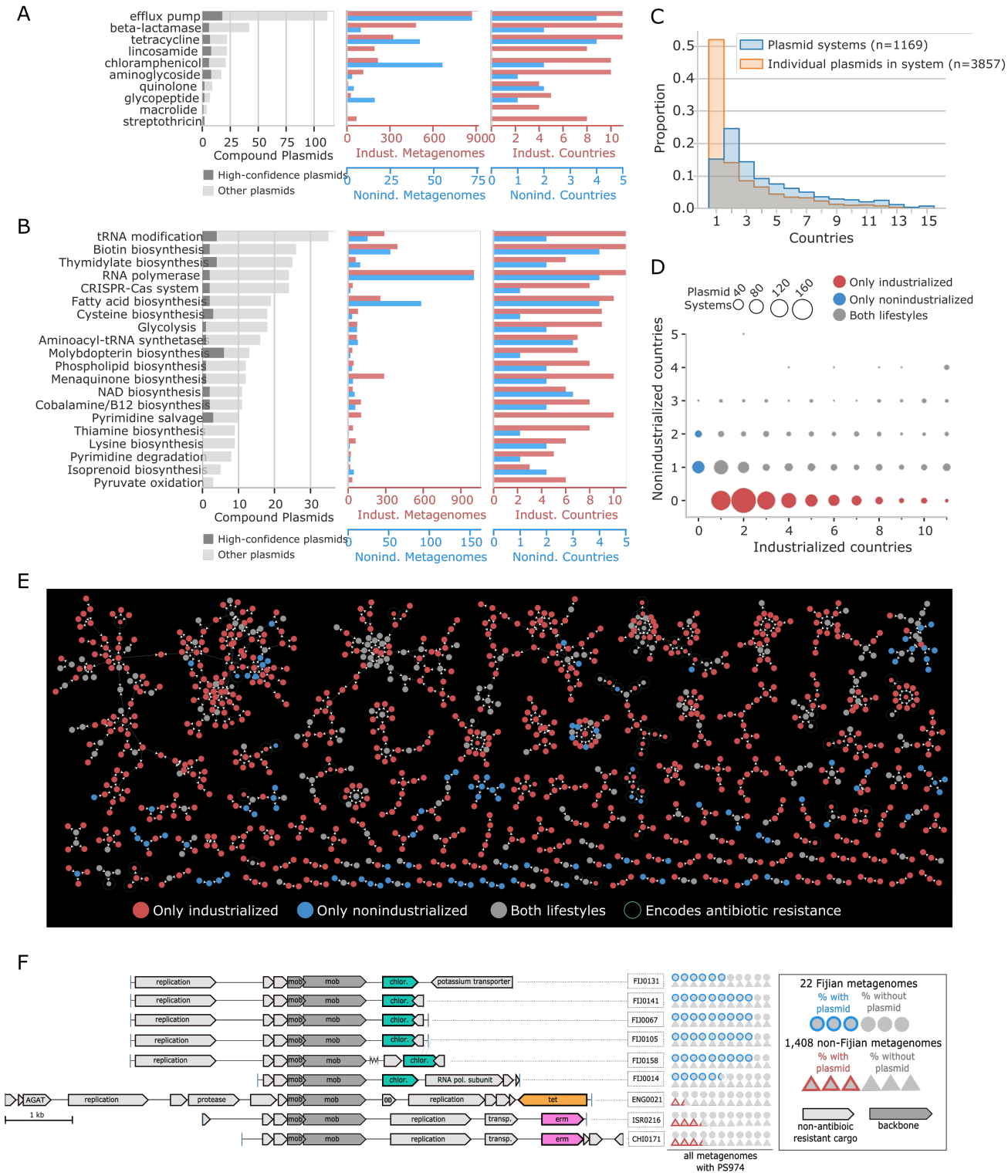
We propose that our algorithmic definition of plasmid systems can be used to study how ecological pressures on plasmids drive the evolution of cargo genes. Plasmid systems are akin to the concept of a genetic 'delivery van' (backbone) with the flexibility to disseminate variable 'packages' (cargo genes) to microbes. The dynamic pool of cargo genes could serve as a means for a plasmid, or its microbial host, to increase fitness in different environmental conditions. To investigate this hypothesis, we compared the evolution of plasmids in industrialized versus non-industrialized countries, an environmental difference that was reflected in plasmid content (Figure 4C). While many plasmids systems were exclusive to one or the other type of country, 396 of them were present in both types (Figure 6D). These global systems provide a unique *in silico* framework to explain environmental differences by variation in plasmid cargo genes.

To demonstrate this framework, we examined antibiotic usage, an extreme environmental difference that is well known to exert selective pressures on microbes, often causing them to maintain plasmids with antibiotic resistance[101–108]. The well-studied nature of antibiotic resistance also provides a testbed to demonstrate that plasmid systems can identify cargo genes under selection. In our data, the evolution of antibiotic resistance in a plasmid system coincided with the ecological variation of compound plasmids in the system. Specifically, we identified 24 high-confidence, compound plasmids that encoded antibiotic resistance in cargo genes and were exclusively present in either non-industrialized or industrialized countries (Figure 6E). Among non-industrialized metagenomes, one of the most common types of antibiotic resistance is chloramphenicol (Figure 6A). For instance, PS974 is highly diverse with 97 non-redundant plasmids; however, this system possesses chloramphenicol resistance (conferred via an acetyltransferase) only in plasmids assembled from Fiji (Figure 6F, Table S5). When we searched for these resistance plasmids across the global set of 1,430 metagenomes that contain PS974, we found them in 19/22 Fijian metagenomes but only in 1/1,408 non-Fijian metagenomes (p=1.1 x 10^{-13}, Fisher's exact test) (Figure 6F, pictogram). Chloramphenicol is routinely prescribed in Fiji to treat eye infections, central nervous system infections, periodontitis, shigellosis, typhoid and paratyphoid fevers, and diabetic foot infections, but it is rarely used in North America and Europe[109–112]. Thus, chloramphenicol resistance in this system likely reflects the increased exposure of Fijians to this antibiotic. While we observed that chloramphenicol-resistant plasmids appeared specific to Fiji, more extensive sampling may reveal the presence of these plasmids in other non-industrialized countries that also have high usage of this antibiotic.

585 Besides chloramphenicol resistance, PS974 also contained non-Fijian compound plasmids that carry tetracycline resistance (171/1,408 metagenomes) or erythromycin resistance (429/1,408 metagenomes) (Figure 6F). In an attempt to find an alternative explanation for the distribution of resistance plasmids in this system, we revisited our earlier question "Can plasmid ecology be simply explained by taxonomy?". By searching for these plasmids among known sequences in
590 NCBI, we determined that possible microbial hosts include Firmicutes, such as *Blautia hydrogenotrophica*. However, none of these hosts nor any other microbial taxon had a similar ecological distribution as any of the resistance plasmids (highest Jaccard index=0.37 across all plasmid-taxon comparisons). These results suggest that compound plasmids in systems have acquired antibiotic resistance to respond to lifestyle-specific usage of antibiotics.

595

While the connection between antibiotic usage and resistance is expected given previous studies[101–108], plasmid systems in general can be used to determine if cargo genes are under selection, even when the functions of those genes or the environmental pressures driving the selection are unknown. For example, the cargo gene encoding tRNA modification in system
600 PS1110 may provide some immunoevasive function (Figure S10), but there was not a clear geographic or lifestyle-specific association with this function. This is an example where we know the cargo function but not the environmental pressure and motivates collecting additional data about the environment. Overall, our work provides a computational roadmap for generating new hypotheses about plasmid evolution on an omics scale.



605

Figure 6. Functional and ecological variation of plasmid systems. (A-B) The number of compound plasmids that encode antibiotic resistance (A) and COG pathways (B) in cargo genes. Also shown are the numbers of metagenomes and countries that contain those plasmids. In B, we show the 20 COG pathways that have the highest number of compound plasmids. To avoid redundancy with A, we exclude COG

610 pathways that occur in cargo genes encoding antibiotic resistance. **(C)** Prevalence of plasmid systems versus the individual plasmids in those systems. **(D)** Distribution of plasmid systems based on the number of industrialized and non-industrialized countries they are found in. **(E)** Recoloring of the network of plasmid systems shown in Figure 5C. Colors indicate whether a plasmid occurred in only industrialized, only non-industrialized, or both types of countries. A green ring indicates a plasmid encoding antibiotic resistance.
615 **(F)** Compound plasmids from PS974 that encode for resistance to chloramphenicol (chlor), tetracycline (tet), or erythromycin (erm). 6/9 plasmids are circular. Dark gray genes are the backbone; light gray are cargo not related to antibiotic resistance. AGAT=aminoglycoside adenylyltransferase. OD=Oxaloacetate decarboxylase. PS974 is found in 22 Fijian and 1,408 non-Fijian metagenomes. The pictogram on the right-hand side represents these metagenomes using two shapes: circles (Fijian) and triangles (non-Fijian). For
620 each plasmid, circles are colored blue to represent the proportion of the 22 Fijian-metagenomes that contain the plasmid. Similarly, triangles are colored red to represent the proportion of the 1,408 metagenomes that contain the plasmid.

Discussion

Our work greatly expands the number of known plasmids by mining a global collection of metagenomes using machine learning. This expansion provides the community with a new resource to study fundamental concepts in plasmid biology. While there are many applications of our resource, here we focused on organizing plasmids into cohesive units known as plasmid systems to gain deeper insights into plasmid ecology and evolution. For instance, the diversity captured by our large collection of 1,169 plasmid systems reveals the great extent to which plasmids in complex ecosystems like the human gut are not static entities but actively evolving in response to the environment. By revealing likely determinants of fitness, such as the acquisition of specific antibiotic resistance genes in Fiji as a response to a commonly used antibiotic, plasmid systems serve as a hypothesis generation and testing tool to study forces that drive plasmid evolution and influence the ecology of hosts that carry them. Our study has focused on geographical or lifestyle-based environmental differences, but more generally, our analysis of plasmid systems can be applied to other contexts such as discerning cargo genes that distinguish healthy vs. disease states of the gut microbiome.

During the past few decades, our ability to bypass the limitations of cultivation and study microbial genomes derived from metagenomes has led to key biotechnological insights[113–116]. The malleability of plasmids is a desirable property in bioengineering and often has motivated the repurposing of naturally occurring plasmids into major tools for genetically modifying organisms. In this vein, we propose computational prediction and analysis of plasmids as an attractive approach to expand the toolkit of available plasmids for genetic engineering, particularly if they can be found in isolates that currently lack tools to make them genetically tractable.

PlasX and other plasmid recognition systems[30,31,33,36,37,46,117–119], along with MobMess to characterize plasmid systems, present a roadmap for a detailed characterization of naturally occurring plasmids. Historically, plasmids and other genetic elements have been characterized on the basis of qualitative properties and descriptions. In contrast, PlasX and machine learning approaches provide an “operational definition” of a plasmid that can be universally and objectively applied. As some of our predicted plasmids contain virus-like or ICE-like signatures, our work can be used to study the spectrum of mobile elements that blur traditional labels and complements recent efforts to characterize viruses[120–123] and horizontally transferred elements[124–126].

To expand the scope of our work, we intentionally designed PlasX using a broad collection of reference sequences, so that it can be applied to study any environment and can include additional training sequences to improve accuracy. These methods provide a complementary

approach to frequently used state-of-the-art workflows to study the taxonomic composition or functional potential of environmental or host-associated microbiomes through amplicon sequences or metagenomes. Overall, our findings suggest that high-throughput recognition and characterization of plasmids in microbiome studies are necessary for more complete insights into the ecology of naturally occurring microbial systems.

Methods

Compiling and annotating a reference set of plasmids and chromosomes

We obtained a list of 16,168 plasmids from the March 5, 2019 version of PLSDB[127]. We also downloaded the entire collection of 13,471 complete bacterial genome assemblies from NCBI RefSeq on October 26, 2019, using instructions at https://www.ncbi.nlm.nih.gov/genome/doc/ftpfaq/#allcomplete. The RefSeq assemblies contained 26,376 contigs, of which we discarded 11,350 that are also in PLSDB. The reference set of 16,827 plasmids consisted of 16,168 PLSDB contigs, as well as 659 contigs from the Refseq assemblies that were labeled as 'Plasmid' in the 'Assigned-Molecule-Location/Type' field of the NCBI assembly report. The reference set of chromosomes was the remaining 14,367 Refeq contigs.

To identify and annotate genes in these sequences, we used the program `anvi-run-workflow` with `--workflow contigs` implemented[128] in anvi'o v7.1[129], which uses Snakemake[130] to execute previously defined steps (https://merenlab.org/anvio-workflows/). These steps include first running Prodigal[131] to call genes and then running DIAMOND v2.0[132] and HMMER v3.3[133] on amino acid sequences to determine gene functions against the Cluster of Orthologous Groups of proteins (COGs)[42] and Protein Family Database models (Pfams) v32.0[43], respectively. To minimize noise, we used an e-value cutoff of 10^{-10} for COGs and the default model scores for Pfams.

Modeling *de novo* gene families

We inferred *de novo* gene families by running MMseqs2[134] v10.6d92c on all amino acid sequences in our reference plasmids and chromosomes. First, we ran `mmseqs clusthash` to collapse identical sequences into a non-redundant set for faster execution of the next step; the collapsing was inverted at the end to annotate all genes. Next, we ran `mmseqs cluster` to calculate pairwise alignments and then cluster genes that are aligned above a minimum sequence identity threshold (parameter `--min-seq-id`). We ran this program multiple times with different thresholds (0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.25, 0.2, 0.15, 0.1, 0.05) to infer a wide range of possible families. Families from different thresholds can be redundant, so we merged nested families, i.e. if family X contains all genes in family Y, then we keep X and discard Y. We also discarded any family that contains only one gene. The final result was 1,090,132 families that annotate 162,783,114 genes (a gene may have multiple annotations). This analysis took advantage of MMseqs2's parallelism, taking ~6 hours using 256 CPU cores.

We refer to a *de novo* family as a 'subfamily' if 90% or more of its amino acid sequences are also annotated to a specific COG or Pfam. Note that this definition provides a small tolerance such that a subfamily does not need to be a perfect subset of a COG or Pfam. For the example about Pfam PF10609 (Figures 1F and 1G), we gathered the 253 amino acid sequences annotated to PF10609 and the subfamily mmseqs_5_1535552. We also gathered the 1,391 sequences annotated to PF10609 and the subfamily mmseqs_70_40217271. We collapsed 100% identical sequences to yield a total collection of 142 and 310 sequences from mmseqs_5_1535552 and mmseqs_70_40217271, respectively. We aligned all of these sequences together using muscle v3.8.1551 (default parameters)[135] and then constructed a maximum likelihood phylogenetic tree

using IQ-TREEv2.1.2 (parameters -m TEST -bb 1000 -alrt 1000 -T AUTO)[136]. We then rooted the tree using the midpoint method.

Subtypes and slicing of reference sequences

To group reference sequences into subtypes, we used mash v2.2.2[137] (command `mash dist`, sketch size 100000, kmer size 21) to calculate a distance score of 0 to 1 between every pair of sequences. Next, we created an undirected graph, where sequences are nodes and sequences are connected if their distance is ≤0.1. We defined a 'subtype' as one of the 7,326 connected components in the graph. 3,935 subtypes contained only plasmids; 3,355 subtypes contained only chromosomes; and 36 subtypes contained both plasmids and chromosomes.

We sliced reference sequences into 10kb slices by sliding a 10kb window at 5kb increments. The first window starts at the beginning of the sequence, and the final window stops at the end of the sequence. For instance, a 23kb sequence would be sliced at 0-10kb, 5-15kb, 10-20kb, and 13-23kb. A slice was annotated with any gene that was entirely or partly inside the slice. In total, we generated 10,453,279 slices from the reference chromosomes and 343,246 slices from the reference plasmids.

Assessing model performance in cross-validation

To perform cross-validation, we randomly divided the 10kb slices into four groups. For each fold of cross-validation, three groups formed the training data, and the fourth group formed the testing data. In a naive split, we keep all slices from the same reference sequence together in either training or testing data. In an informed split, we keep all slices from the same subtype together.

We assigned weights to the 10kb slices when calculating precision and recall performance (Figure S1F and 1D). Consider the following notation to represent sequences:

$$S_i \;=\; sequence\ i$$

$$P_u \;=\; the\ set\ of\ plasmid\ sequences\ in\ subtype\ u$$

$$C_u \;=\; the\ set\ of\ chromosome\ sequences\ in\ subtype\ u$$

$$D_i^k \;=\; window\ slice\ k\ of\ sequence\ i$$

And consider the following notation to represent weights:

$$w(D_i^k) \;=\; weight\ of\ window\ slice\ k\ of\ sequence\ i$$

$$w(S_i) = \sum_k w(D_i^k) \;= weight\ of\ sequence\ i$$

$$w(P_u) \;=\; \sum_{S_i \in P_u} w(S_i) = weight\ of\ plasmid\ sequences\ in\ subtype\ u$$

$$w(C_u) = \sum_{S_i \in C_u} w(S_i) \;= weight\ of\ chromosome\ sequences\ in\ subtype\ u$$

We defined two different scenarios for assigning weights. Scenario A satisfies the following conditions:

1.) All slices from the same sequence have the same weight

$$w(D_i^s) = w(D_i^t) \quad \forall s, t$$

2.) The weight of every sequence is equal to 1

$$w(S_i) = 1 \quad \forall i$$

Scenario B satisfies the following conditions:

1.) All slices from the same sequence have the same weight

$$w(D_i^s) = w(D_i^t) \quad \forall s, t$$

2.) All plasmid (or chromosome) sequences in the same subtype have equal weight

$$w(S_i) = w(S_j) \quad \forall\, i, j, u \text{ where } S_i \in P_u \text{ and } S_j \in P_u$$

$$w(S_i) = w(S_j) \quad \forall\, i, j, u \text{ where } S_i \in C_u \text{ and } S_j \in C_u$$

3.) All subtypes have equal weight

$$w(P_u) = w(P_v) \quad \forall u, v$$

$$w(C_u) = w(C_v) \quad \forall u, v$$

4.) The sum of weights across all slices equals the total number of slices

$$\sum_u w(P_u) = \text{total number of plasmid slices } (i.e. 343{,}246)$$

$$\sum_u w(C_u) = \text{total number of chromosome slices } (i.e. 10{,}453{,}279)$$

Each scenario implies a unique assignment of weight values. Scenario A requires that every sequence has the same weight. Importantly, this ensures that long sequences, which have disproportionately more slices, have equal weight as shorter sequences. Scenario B further requires that every subtype has the same weight. Importantly, this ensures that subtypes that contain a disproportionately large number of sequences (e.g. subtypes that represent commonly studied bacteria, such as *Escherichia*, *Salmonella,* and *Klebsiella*) have equal weight as subtypes with fewer sequences.

We evaluated performance under two different cross-validation and weighting scenarios. Figure S1F shows the result of training models using a 'naive' cross-validation split and calculating precision/recall using weights from Scenario A. Figure 1D shows the results of training models using an 'informed' cross-validation split and calculating precision/recall using weights from Scenario B. We calculated precision/recall using the function *sklearn.metrics.precision_recall_curve* from the scikit-learn Python package[138], with the parameter *sample_weight* set to the weights of the slices. We calculated AUCPR with the function *sklearn.metrics.average_precision_score*.

PlasX implementation

We implemented PlasX as a logistic regression using the *SGDClassifier* class from scikit-learn[138]. Regardless of how we evaluated PlasX, we always trained it with weights defined by Scenario B and based on only slices in the training data. To implement elastic net regularization, we performed a grid search of hyperparameters, with the regularization parameter *alpha* ranging from 10^{-8} to 10^{-3} in multiplicative increments of $\sqrt{10}$ and the parameter *l1_ratio* being 0.0, 0.25, 0.5, 0.75, or 1.0. For each evaluation scenario, we selected the hyperparameters that produced the best performance. We used the best hyperparameters from the 'informed' cross-validation and the weights defined by Scenario B (*alpha*=3.16×10^{-6}, *l1_ratio*=0.0) to retrain PlasX on all 10kb slices and create the final model that we used to predict plasmids from metagenomes.

Execution of other plasmid prediction tools

We downloaded PlasClass[31] from https://github.com/Shamir-Lab/PlasClass (v0.1.0-2-gb80a4f4). We downloaded PPR-Meta[45] from https://github.com/zhenchengfang/PPR-Meta (v1.0-14-gab99c91). We downloaded Platon[46] from https://github.com/oschwengers/platon, and then modified the code to more efficiently parallelize across many CPUs (modifications at https://github.com/michaelkyu/platon). We used Platon's RDS score as its final prediction score, ignoring whether it found other features like conjugation and replication genes.

To ensure a fair comparison of models in cross-validation (Figure S1F and 1D), we retrained PlasClass and Platon using the same training sequences as we used for PlasX in each cross-validation fold. We trained PlasClass on 10kb slices, and we trained Platon on the entire non-sliced sequences. We did not train PlasClass and Platon with sequence weights because they don't take in weights as input, but we did calculate precision and recall with weights. PPR-Meta[45] and Deeplasmid[35] do not provide software interfaces for retraining new models, so we ran the pretrained versions of these models that were published in their original studies (and thus were trained on different sequence datasets).

We downloaded the four sequence versions of the Wolbachia plasmid pWCP from https://doi.org/10.6084/m9.figshare.6380015 (Supplementary Table S8). We made predictions of pWCP using the original pretrained and published versions of PlasClass, Platon, PPR-Meta, and Deeplasmid, and we ran the final PlasX model that was trained on all 10kb slices.

We ran Deeplasmid using the Docker image of the CPU implementation, following instructions at https://github.com/wandreopoulos/deeplasmid (version *sha256:10809927e2c8a14cf86231801b804b0bd4bddf600821d17fd8b7e41a15c562c0*). While we were able to run Deeplasmid on the Wolbachia plasmid pWCP, it was prohibitively slow to run on the entire set of 10kb slices used for cross-validation evaluation. In particular, we found that Deeplasmid running on a MacOS laptop takes ~3 hours for 1,000 slices, so we estimated it would take ~3.7 years to run on all slices. While the GPU implementation of Deeplasmid might be able to run faster, we were unable to execute its prebuilt Docker image (version sha256:f3a22993fb765a7f9678b174245b64976e7e52a4dce85570060900b794af5e43). We suspect that this image is incompatible with modern machine setups, like ours, because Deeplasmid depends on software that is several years old. For example, it requires the CNTK library, for which development was abandoned over 3 years ago (https://docs.microsoft.com/en-us/cognitive-toolkit/releasenotes/cntk_2_7_release_notes). We were also unable to build a new Docker image to run the GPU implementation, despite attempts to modify the Docker build file (see the issue we raised at https://github.com/wandreopoulos/deeplasmid/issues/3).

Predicting plasmids from metagenomic assemblies

We downloaded fastq files for 1,782 short-read and paired-end metagenomes from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) using the program `fastq-dump`. The countries represented are Austria[139], Australia (https://www.ncbi.nlm.nih.gov/bioproject/PRJEB6092), Bangladesh[140], Canada[141], China[142,143], Denmark[144], England[145], Ethiopia[68], Fiji[126], Israel[146], Italy[147], Madagascar[68], Mongolia[148], Spain[95], Tanzania[147] and USA[93,149]. Some samples were sequenced multiple times (i.e. multiple records in SRA), in which case we concatenated the fastq files together. We have separated these multiple accessions using the delimiter '|'. We labeled Tanzania, Ethiopia, Bangladesh, Madagascar, and Fiji as non-industrialized and the other countries as industrialized.

All steps of quality filtering, metagenomic assembly, read recruitment and profiling were automated using snakemake[130] workflows in anvi'o[150]. The `illumina-utils`[151] commands `iu-gen-configs` and `iu-filter-quality-minoche` with the flag `--ignore-deflines` were used to quality filter the raw paired-end reads. Each metagenome was assembled individually using IDBA_UD[152] with default settings, except the flag `--min_contig 1000`.

We annotated COGs and Pfams in all assembled contigs using the same procedure as the reference plasmids and chromosomes. To annotate *de novo* families, we first used `mmseqs result2profile` (default parameters) to represent the sequence conservation in each *de novo* family as a profile. We then used `mmseqs search` (default parameters) to search for profiles across all genes. We kept hits where the alignment coverage was ≥80% of both the gene and the profile and where the alignment identity was at least ≥X-0.05 where X is the minimum identity threshold used to originally construct the family (parameter --min-seq-id). For example, if a family was constructed using an identity threshold of 0.8, then we kept hits with an identity ≥0.75. Using these gene annotations, we ran PlasX to assign a score to every contig. We kept contigs intact, rather than slicing them into 10kb windows. Contigs with score >0.5 were classified as plasmids.

Detection and circularity of plasmids across metagenomes

We recruited short reads from our collection of metagenomes using Bowtie2 v.2.0.5[153]. We used the snakemake workflows in anvi'o to automate execution of bowtie and post-processing to calculate 'detection', i.e. the proportion of a sequence that is covered by at least one read. We ran bowtie2 using the three following combinations of parameters and input files.

First, to identify circular contigs, we recruited each metagenome's reads to a fasta file that contained only the contigs assembled from that metagenome. For computational efficiency, we ran bowtie2 with its default behavior to align every read at most once. We then analyzed the orientation of paired-end reads (Figure 2B). During assembly, circular sequences are broken by an artificial breakpoint to represent them as linear contigs. Consequently, DNA sequencing that occurred across this breakpoint will produce paired-end reads that align in a reverse-forward orientation to the ends of the contig. In contrast, if a sequence is not circular, then all paired-end reads are expected to align in a forward-reverse orientation. To illustrate this intuition, suppose the upstream read of a paired-end maps to positions 200-300 of a contig and the downstream read maps to 500-600. If the upstream read maps with a reverse complement strandedness (i.e. 'reverse') and the downstream read maps with the same strandedness as the way the contig is written (i.e. 'forward'), then the paired-end is in a reverse-forward orientation. In other words, if the contig is written 5'-to-3', then the upstream read maps 3'-to-5' and the downstream read maps 5'-to-3'. Inversely, the paired-end is in a forward-reverse orientation if the upstream read maps 5'-to-3' and the downstream read maps 3'-to-5'. Next, we defined the gap (or insert) size of a paired-end to be the distance between the farthest (or closest) aligned positions between its two reads. In our example, the gap size is 600-200=400 and the insert size is 500-300=200. Let D be the contig's length minus three times the median insert size of all forward-reverse paired-ends that are aligned to the contig. Finally, we label a contig as circular if (1) its detection was ≥0.95 and (2) it had at least one reverse-forward paired-end with a gap greater than or equal to D. This approach of examining reverse-forward paired-ends was inspired by[154].

Second, to study the ecological distribution of all plasmids and plasmid systems at the same time, we recruited each metagenome's reads to a fasta file that contained either the non-

redundant set of 68,350 predicted plasmids or the non-redundant set of 11,121 reference plasmids. For computational efficiency, every read was aligned at most once (i.e. the default behavior of bowtie). We designated a plasmid as being present in a metagenome if its detection was ≥0.95. To compare metagenomes based on their plasmid content in Figure 4 and Figure S4, we ran UMAP v0.5.1[75] with parameters `n_neighbors=30, n_components=2, min_dist=0.15, metric='jaccard', random_state=1`. The heatmaps in Figure 4 were generated using the `heatmap.2` package in R, with agglomerative clustering using median linkage on Euclidean distances.

Third, to study the specific plasmids from PS974 and PS1110 that were shown in Figure 6F and Figure S10 (see Table S5 for contig names), we ran bowtie2 on each sequence separately. This setup allowed every read to align to potentially multiple sequences, resulting in a more complete estimation of which metagenomes contained a plasmid. For the backbone sequences of these systems, we designated them as present in a metagenome if their detection was ≥0.95. For compound plasmids in PS1110 that encoded a chloramphenicol resistance gene, we designated them as present in a metagenome if they satisfied an additional criterion that ≥0.95 of the resistance gene was covered by at least one read.

Estimation of microbial taxonomy in metagenomes

We estimated taxonomic abundances in every metagenome by running kraken2 v2.1.2 with its standard database (https://github.com/DerrickWood/kraken2)[155] and then refined the abundances using bracken v2.5 (https://github.com/jenniferlu717/Bracken)[156] with database parameters `-k 35 -l 96`. We ran bracken (parameter `-r 96`) a separate time for every taxonomic rank: S1 (subspecies/strain), S (species), G (genus), F (family), O (order), C (class), P (phylum), D (domain). The output of this analysis is a count of how many reads originated from each taxon.

To compare the metagenomic presence/absence of plasmids versus taxa, we calculated M_P, the set of metagenomes where a plasmid P is detected at ≥95%, and M_T^r, the set of metagenomes in which at least r reads originated from taxon T. For each plasmid, we attempted to find the best explanation of its ecological distribution by comparing the plasmid to every taxon using the Jaccard index, and by scanning many possible read thresholds. More exactly, we used the following formula to represent the best possible explanation of a plasmid's ecological distribution:

$$\max_T \max_r Jaccard(P, T; r)$$

where

$$Jaccard(P, T; r) = \frac{|M_P \cap M_T^r|}{|M_P \cup M_T^r|}$$

We evaluated 29 values for the threshold r, ranging from 1 read to 10 million reads in multiplicative increments of $10^{1/4}$. We ignored plasmids that were present in less than 5 metagenomes, i.e. $|M_P|<5$, because it was likely that these plasmids would have a high Jaccard similarity to some taxon by chance. For instance, we observed that many pairs of plasmids and taxa occur in exactly one and the same metagenome, and thus they have a Jaccard index of 1.

To compare continuous-valued abundances, we defined a plasmid's abundance in a metagenome as the sum of coverage values across all sequence positions divided by sequence length, and we defined a taxon's abundance as the number of reads originating from the taxon. If

a plasmid had detection of <95%, then we set its abundance to 0. If a taxon had less than 1000 reads, then we set its abundance to 0. We ignored plasmids and taxa that had non-zero abundances in less than 5 metagenomes. For every pair of plasmid and taxon, we estimated the Pearson correlation between their abundance levels across metagenomes using FastSpar v1.0.0 (https://github.com/scwatts/fastspar)[157], which is an improved implementation of the SparCC method[76]. This method accounts for the compositional nature of the data—in which abundances reflect relative instead of absolute quantities—by assuming that the amount of correlations in a data is sparse. We ran FastSpar on the non-redundant set of predicted plasmids, and ran it separately on the non-redundant set of reference plasmids.

Additional validation and annotations of plasmids

To determine if a predicted plasmid has canonical plasmid features, we ran MOB-suite[38]. This tool searches a sequence for known examples of four types of features: plasmid replicon (e.g. replication genes), relaxase, mating pair formation, and origin of transfer. We installed MOB-suite v3.0.1 using pip, in an Anaconda Python environment that has mash v2.2. We ran the MOB_typer subroutine (command `mob_typer`) using default parameters and followed the execution instructions at https://github.com/phac-nml/mob-suite.

To determine if a predictive plasmid is a novel sequence, plasmids were blasted against NCBI using the *blast* package (v2.9.0, installed from bioconda Anaconda repository). On October 13, 2021, we downloaded version 5 of the NCBI databases *non-redundant nucleotide* (*nt*), *ref_prok_rep_genomes*, *ref_viroids_rep_genomes*, and *ref_viruses_rep_genomes*, and then integrated them into a single database using the `blastdb_aliastool` command. We then searched every predicted plasmids against this combined database, using the `blastn` tool with the '-task megablast' parameter for efficient searching. For each plasmid, we examined all matching NCBI sequences (called 'subjects') and chose the one with the highest 'qcovs' (query coverage per subject), which represents the fraction of the plasmid sequence that is covered by all high-scoring segment pairs (HSP). Tiebreaking was done by sorting subjects by the maximum bitscore of the HSPs. If the qcovs of the best matching sequence was ≥90%, then we considered the predicted plasmid as found in NCBI and further categorized the matching sequence by searching for the keywords 'plasmid', 'virus', 'chromosome' (in that order, disregarding capitalization) in its NCBI description. For example, if the description of the matching sequence contained the word 'plasmid', then we said the predicted plasmid matched a known plasmid on NCBI. Similarly, if the description contained 'chromosome' but not 'plasmid' nor 'virus', then we said that the predicted plasmid matched a known chromosome on NCBI. If the qcovs of the best sequence was <90%, then we labeled the predicted plasmid as not found in NCBI.

To visualize pFIJ1037_1 (Figure 3A) and pENG0187_1 (Figure S3A), we manually imported COG functions into the plasmid maps produced by snapgene (Insightful Science; snapgene.com). We manually curated functions in genes without COGs using NCBI BLASTx.

Keyword analysis of COGs and Pfams for plasmid functions

We labeled COGs and Pfams as being a plasmid-associated function (Figure 1E) if its database description contains any of following keywords as a substring: *'plasmid', 'toxin', 'replicat', 'integrase', 'transpos', 'recombinase', 'resolvase', 'relaxase', 'recombination', 'partitioning', 'mobilis', 'mobiliz', 'type iv', 'conjugal', 'conjugat', 'segregat', 'MobA', 'ParA', 'ParB', 'BcsQ'*. We

labeled backbone and cargo genes as being related to plasmid replication, transfer, or maintenance if they were annotated to any plasmid-associated COG or Pfam.

To determine if a predicted plasmid is 'keyword-recognizable' (Figure 2C), we searched the plasmids for COGs and Pfams using a more restricted set of keywords (just "plasmid" and "conjugation") instead of the keywords above.

MobMess algorithm to dereplicate plasmids and discover plasmid systems (see Figure S8)

The MobMess algorithm consists of several steps to study the evolution of plasmids. MobMess first performs an all-vs-all pairwise alignment of sequences using the MUMmer alignment package (v4.0.0rc1)[158]. All sequences are placed into a single fasta file and then aligned with `nucmer` (parameters `--maxmatch --minmatch=16`) to calculate local alignment blocks. Alignments are specified asymmetrically such that one sequence is designated as the query *q* and the other is the reference *r*. For every *q* and *r*, the alignment blocks calculated by `nucmer` are written to a separate file, and then a subset of blocks is identified using `delta-filter` (parameters `-q -r`) to create a one-to-one alignment.

Next, MobMess constructs a directed graph *G* where vertices are sequences and edges represent the containment of one sequence within another (Figure S8A). Formally, consider a query *q* and reference *r*. Let |*q*| be the length of *q*. For the ith alignment block between *q* and *r*, let s^i , e^i , and δ^i be the start position in *q*, end position in *q*, and number of alignment mismatches and indels, respectively. The following values summarize the information across all alignment blocks between *q* and *r*.

Sum of block lengths $$L = \sum_i e^i - s^i$$

Number of mismatches and indels $$E = \sum_i \delta^i$$

Proportion of query positions covered $$C = \sum_{j=1}^{|q|} \begin{cases} 1/|q| & \text{if } \exists i \text{ such that } s^i \leq j \leq e^i \\ 0 & \text{otherwise} \end{cases}$$

Local sequence identity $$I_{local} = (L - E)/L$$

Global sequence identity $$I_{global} = I_{local} * C$$

MobMess creates a directed edge (*q*,*r*) in *G* if $I_{local} \geq 0.9$ and $C \geq 0.9$.

MobMess clusters sequences according to strongly connected components in *G*, calculated with igraph v0.8.2[159] in Python. That is, two sequences *x* and *y* are placed in the same cluster if there exists a directed path from *x* to *y* and another from *y* to *x* in *G*. Intuitively, a cluster represents a set of sequences that are nearly identical to each other across nearly their entire lengths. MobMess then reduces *G* to another graph *H*, called the condensation graph, by contracting every cluster of sequences into a single vertex. A directed edge (*u*,*v*) exists in *H* if and only if there are sequences $x \in u$ and $y \in v$ where edge *(x,y)* exists in *G*. Note that *H* does not have any cycles. As proof by contradiction, if there were a cycle of clusters, then those clusters

would have been in the same strongly connected component in *G* and hence would have been merged into a single, larger cluster.

995 MobMess labels every cluster in *H* as one of the three following types: (1) a 'backbone cluster' if it has an outgoing edge and at least one of its member sequences is circular, (2) a 'fragment cluster' if it has an outgoing edge but none of its member sequences are circular, or (3) a 'maximal cluster' if it does not have any outgoing edges. Intuitively, a maximal cluster represents the longest version of a plasmid observed in the data. In contrast, a fragment cluster is given this
1000 name because it may be the assembly fragment of a maximal cluster.

To dereplicate sequences, MobMess discards all fragment clusters and then chooses a representative sequence from every maximal and backbone cluster. A cluster's representative is the sequence with the highest global sequence identity (I_{global}), averaged across the set of alignments where that sequence is the reference and other sequences in the same cluster are
1005 the queries.

MobMess defines a plasmid system as a specific backbone cluster together with its 'compound' clusters, which are the set of non-fragment clusters connected to the backbone in *H*. Thus, there is a one-to-one correspondence between backbone clusters and plasmid systems. Note that systems can be nested within each other, because backbone clusters can be connected
1010 to each other in *H*. Thus, a backbone cluster can be *the backbone* that forms a given plasmid system, and at the same time, it can also be a compound cluster with respect to an even smaller backbone.

We ran MobMess to analyze the 226,164 predicted plasmid sequences, yielding 132,616 clusters. Of these, 68,350 were non-fragment clusters, containing 100,719 contigs. We ran
1015 MobMess separately on the 16,827 reference plasmid sequences, yielding 11,121 clusters. We assumed that all reference plasmids were circular, and thus there were no fragment clusters. We visualized networks with Cytoscape v3.8[83] and laid nodes out using the prefuse directed force layout[84]. While we have focused on plasmids, MobMess could be applied to dereplicate and organize other mobile genetic elements into systems.

1020

Classification of cargo and backbone genes

We classified genes in a plasmid system as being backbone or cargo. We defined backbone genes as any gene that is encoded in any backbone plasmid of the system. We defined cargo genes as any gene that (1) is encoded in any compound plasmid of the system and (2) does not
1025 share any *de novo* family annotations with the backbone genes. For this second condition, we used the 1,090,132 *de novo* families that we constructed from reference plasmids and chromosomes in order to train PlasX. We also constructed and used an additional set of *de novo* families by running the command mmseqs (--min-seq-id 0.05) on the genes from all plasmid sequences in this study (16,827 reference and 226,194 predicted plasmids). This additional
1030 analysis allowed us to analyze gene families that might be absent in reference sequences but are conserved in predicted plasmids.

Mutual exclusivity of plasmid systems

To measure the extent of mutual exclusivity in a plasmid system *S*, we defined two sets of
1035 metagenomes. M_S^{any} is the set of metagenomes that contains *S*, i.e. where one or more plasmids in *S* has detection at ≥0.95. M_S^{solo} is the set of metagenomes where one and only one plasmid in

S has detection at ≥0.95. Then, we calculated the mutual exclusivity score E_S as the ratio $|M_s^{solo}|$ / $|M_s^{any}|$. This score is similar to a test statistic used by methods, such as CoMEt[160], for studying mutual exclusivity of gene alterations in cancer.

We observed that if a compound plasmid was present in a metagenome (detection ≥0.95), then its backbone plasmid was often also present. This could be due to (1) both plasmids being present as separate entities in the same microbiome, or (2) a read recruitment artifact arising from the sequence similarity between the plasmids. To minimize artifacts, we assumed the second scenario is always happening, and we corrected for it by assuming that a backbone plasmid is absent from a metagenome (regardless of its detection) whenever any compound plasmids are present in the metagenome. To formalize this correction procedure, let P_m be the set of plasmids in system S with detection ≥0.95 in metagenome *m*. We created the induced subgraph H_m that is formed by subsetting the vertices of P_m in the plasmid similarity graph *H* (see MobMess section of the Methods). Then, we defined $P'_m \subseteq P_m$ as the subset of plasmids that are maximal with respect to H_m (i.e. they don't have outgoing edges in H_m). We used P'_m, rather than P_m, in order to calculate M_s^{any}, M_s^{solo}, and E_S.

Identification of antibiotic resistance genes

We annotated antibiotic resistance genes using two databases. First, we searched against a database of resistance protein family HMMs from Resfams (v1.2, dated 2015-01-27, 'Core' database at http://www.dantaslab.org/resfams)[161]. We used `anvi-run-hmms` from anvi'o[129] to automate running `hmmsearch` from HMMER 3.3.2[133] and apply an e-value cutoff of 10^{-10}. Second, we ran rgi (v5.2.0, https://github.com/arpcard/rgi) to search for similarity in the CARD database of resistance genes[162]. We removed CARD hits that were labeled as 'Loose' and kept those labeled as 'Perfect' or 'Strict'. We removed any Resfams or CARD hits that contained the keywords 'transcription', 'regulat', 'modulat' in their database description, to avoid cases (e.g. TetR protein) where the hit is a gene that regulates the expression of another resistance gene but doesn't itself perform the molecular process that confers resistance. We categorized hits into major antibiotic resistance classes by searching for the following keywords in their functional descriptions: 'lincosamide', 'macrolide', 'erythromycin', 'chloramphenicol', 'aminoglycoside', 'streptothricin', 'glycopeptide', 'efflux pump', 'beta-lactamase', 'nitroimidazole', 'tetraycyline', 'quinolone', 'sulfonamide'. Additionally, we searched the extra keywords 'Van' and 'VanZ' to identify glycopeptide resistance; 'efflux', 'permease', and 'pump' to identify efflux pumps; and 'TetX' to identify tetracycline resistance.

High molecular weight (HMW) DNA extraction, long-read sequencing, and determination of circularity through long-reads

We employed a long-read sequencing strategy on two *Bacteroides fragilis* cultivars from two patients (p-214 and n-216 previously described in Vineis et al[65]). We extracted total genomic HMW DNA by one of two methods. For *B. fragilis* p-214, we used the Qiagen Genomic Tip 20/G procedure (also known as Method #4/GT) as previously described[163] on a 10 mL overnight BHIS broth culture. For *B. fragilis* n-216, we used a Phenol Chloroform protocol[163] on 25 mL overnight BHIS broth cultures. Libraries were prepared with the Rapid Barcoding Kit (SQK-RBK004) and the standard protocols from Oxford Nanopore Technologies (ONT) with a few modifications. For *B. fragilis* p-214, DNA fragmentation was performed on 6 ug DNA using 5 passes through a 22G

needle in a 30 µL volume. The gDNA input was 1.5 µg (Table S6), based on sample availability in a 7.5 µL volume, with 2.5 µL Fragmentation mix added. We sequenced for 72 hours using a single R9.4/FLO-MIN106 flow cell (ONT). For *B. fragilis* n-216, DNA fragmentation was performed on 10 ug DNA using 10 passes through a 22G needle in a 250 µL volume. The gDNA input was 0.32 to 0.44 µg, based on sample availability in an 8.5 µL volume, with 1.5 µL Fragmentation mix was added per sample. We sequenced for 72 hours using a single R9.4/FLO-MIN106 flow cell. We used Guppy (v4.0.15) for all post-run base calling, sample de-multiplexing and the conversion of raw FAST5 to FASTQ files.

To determine circularity, we used BLAST to align the long reads with a minimum quality score of 7 to our predicted plasmid sequences. During assembly, all DNA short reads are assembled as linear sequences even if they are circular elements. Circular elements have an artificial breakpoint to represent them as linear sequences, and this breakpoint can happen anywhere on the sequence depending on the assembly method. We tested for the presence of an artificially introduced breakpoint by aligning 500 long reads and then visualizing these alignments on the sequence as if it were assumed to be a circular element (Figure S3). If indeed the sequence is circular, the long reads would overlap each other and “wrap around” the entire circumference of the sequence. In other words, all nucleotide positions of the sequence would be covered by at least one read and there would also exist a read that spans the breakpoint by aligning to both sides of the breakpoint. This property ensures the breakpoint is artificial, and hence the sequence is a circular element. Inversely, this property does not hold when the breakpoint is not artificial (i.e. the sequence is actually an assembly fragment or linear element).

Transfer of predicted plasmid between microbial populations

In duplicate, we streaked *B. fragilis* 214 (donor, erythromycin resistant due to pFIJ0137_1) and *B. fragilis* 638R (recipient, rifampicin resistant) onto plates with brain-heart infusion agar supplemented with hemin and vitamin K (BHIS) and incubated them in 5 mL BHIS media anaerobically at 37℃ for 20 hours. To mate the donor to the recipient, 250 µL of donor cells were pelleted in a centrifuge at 5,000x gravity. We discarded the supernatant and resuspended the donor in 1 mL of the recipient culture. Again, cells were pelleted at 5,000x gravity, then resuspended in 25 µL of BHIS media. Cells were spotted onto BHIS agar plates and incubated anaerobically for 24 hours. After incubation, cells were resuspended in 1 mL BHIS. 250 µL of this suspension was plated onto BHIS plates containing 8 µg/mL rifampicin and 25 µg/mL erythromycin to select for *B. fragilis* 638R recipients of pFIJ0137_1. Duplicate plates each had approximately 300 colonies. Plating the donor or recipient alone resulted in zero colonies, confirming the transformants were not spontaneous mutants to either antibiotic. Two transformant colonies were restreaked onto fresh BHIS plates containing 8 ug/mL rifampicin and 25 µg/mL erythromycin.

Short-read sequencing of isolate genomes and confirmation of plasmid transfer

We grew 20-hour cultures of *B. fragilis* 214 donor, naive *B. fragilis* 638R, and *B. fragilis* 638R transconjugants containing pFIJ0137_1. Using the QIAseq FX DNA library kit (Qiagen), libraries of these strains were prepared with 100 ng of genomic DNA. DNA was fragmented enzymatically into smaller fragments and desired insert size was achieved by adjusting fragmentation conditions. Fragmented DNA was end repaired and ‘A’s were added to the 3’ ends to stage inserts

for ligation. During the ligation step, Illumina compatible Unique Dual Index (UDI) adapters were added to the inserts and the prepared library was PCR amplified. Amplified libraries were cleaned up, and QC was performed using a tapestation. Libraries were sequenced on Illumina MiSeq platform using v2 cassette to generate 2x250bp reads. To confirm the transfer of pFIJ0137_1, we individually recruited reads from the *B. fragilis* 214 donor, naive *B. fragilis* 638R, and *B. fragilis* 638R transconjugants to the pFIJ0137_1 reference sequence. We used anvi'o to create contigs and profile databases (as described above) and visualized these results with the command `anvi-interactive`. We independently confirmed the presence of pFIJ0137_1 by assembling genomes using SPAdes[164] with default parameters.

References

1. Frost, L. S., Leplae, R., Summers, A. O. & Toussaint, A. Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.* **3**, 722–732 (2005).
2. Kumar, D. *et al.* The selfish yeast plasmid utilizes the condensin complex and condensed chromatin for faithful partitioning. *PLoS Genet.* **17**, e1009660 (2021).
3. Kazlauskas, D., Varsani, A., Koonin, E. V. & Krupovic, M. Multiple origins of prokaryotic and eukaryotic single-stranded DNA viruses from bacterial and archaeal plasmids. *Nat. Commun.* **10**, 1–12 (2019).
4. del Solar, G., Giraldo, R., Ruiz-Echevarría, M. J., Espinosa, M. & Díaz-Orejas, R. Replication and Control of Circular Bacterial Plasmids. *Microbiol. Mol. Biol. Rev.* **62**, 434 (1998).
5. Khan, S. A. Rolling-circle replication of bacterial plasmids. *Microbiology and molecular biology reviews : MMBR* vol. 61 442–455 (1997).
6. Lilly, J. & Camps, M. Mechanisms of Theta Plasmid Replication. *Microbiol Spectr* **3**, PLAS–0029–2014 (2015).
7. Thomas, C. M. *Horizontal Gene Pool: Bacterial Plasmids and Gene Spread*. (CRC Press, 2003).
8. Summers, D. *The Biology of Plasmids*. (1993).
9. Jacob, A. E. & Hobbs, S. J. Conjugal transfer of plasmid-borne multiple antibiotic resistance in Streptococcus faecalis var. zymogenes. *J. Bacteriol.* **117**, 360–372 (1974).
10. Poyart-Salmeron, C., Carlier, C., Trieu-Cuot, P., Courtieu, A. L. & Courvalin, P. Transferable plasmid-mediated antibiotic resistance in Listeria monocytogenes. *Lancet* **335**, 1422–1426 (1990).
11. Lan, R., Stevenson, G. & Reeves, P. R. Comparison of Two Major Forms of the Shigella Virulence Plasmid pINV: Positive Selection Is a Major Force Driving the Divergence. *Infect. Immun.* **71**, 6298 (2003).
12. Meletzus, D., Bermphol, A., Dreier, J. & Eichenlaub, R. Evidence for plasmid-encoded virulence factors in the phytopathogenic bacterium Clavibacter michiganensis subsp. michiganensis NCPPB382. *J. Bacteriol.* **175**, 2131–2136 (1993).
13. Sen, D. *et al.* Broad-host-range plasmids from agricultural soils have IncP-1 backbones with diverse accessory genes. *Appl. Environ. Microbiol.* **77**, 7975–7983 (2011).
14. Holt, K. E. *et al.* Multidrug-resistant Salmonella enterica serovar paratyphi A harbors IncHI1 plasmids similar to those found in serovar typhi. *J. Bacteriol.* **189**, 4257–4264 (2007).

15. Fernandez-Lopez, R., Redondo, S., Garcillan-Barcia, M. P. & de la Cruz, F. Towards a taxonomy of conjugative plasmids. *Curr. Opin. Microbiol.* **38**, 106–113 (2017).
16. Oliva, M. *et al.* Antimicrobial resistance gene shuffling and a three-element mobilisation system in the monophasic Salmonella typhimurium strain ST1030. *Plasmid* vol. 111 102532 (2020).
17. Orlek, A. *et al.* Plasmid Classification in an Era of Whole-Genome Sequencing: Application in Studies of Antibiotic Resistance Epidemiology. *Front. Microbiol.* **0**, (2017).
18. Norberg, P., Bergström, M., Jethava, V., Dubhashi, D. & Hermansson, M. The IncP-1 plasmid backbone adapts to different host bacterial species and evolves through homologous recombination. *Nat. Commun.* **2**, 268 (2011).
19. Heuer, H. & Smalla, K. Plasmids foster diversification and adaptation of bacterial populations in soil. *FEMS Microbiol. Rev.* **36**, 1083–1104 (2012).
20. Sota, M. *et al.* Shifts in the host range of a promiscuous plasmid through parallel evolution of its replication initiation protein. *ISME J.* **4**, 1568–1580 (2010).
21. Handelsman, J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.* **68**, 669–685 (2004).
22. Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A. M. & Banfield, J. F. Accurate and complete genomes from metagenomes. *Genome Res.* **30**, 315–333 (2020).
23. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
24. Manor, O. *et al.* Health and disease markers correlate with gut microbiome composition across thousands of people. *Nat. Commun.* **11**, 1–12 (2020).
25. de Vos, W. M., Tilg, H., Van Hul, M. & Cani, P. D. Gut microbiome and health: mechanistic insights. *Gut* **71**, 1020–1032 (2022).
26. Smalla, K., Jechalke, S. & Top, E. M. Plasmid detection, characterization and ecology. *Microbiology spectrum* **3**, (2015).
27. Jones, B. V. & Marchesi, J. R. Transposon-aided capture (TRACA) of plasmids resident in the human gut mobile metagenome. *Nat. Methods* **4**, 55–61 (2007).
28. Delaney, S., Murphy, R. & Walsh, F. A Comparison of Methods for the Extraction of Plasmids Capable of Conferring Antibiotic Resistance in a Human Pathogen From Complex Broiler Cecal Samples. *Front. Microbiol.* **9**, 1731 (2018).
29. Brown Kav, A. *et al.* Insights into the bovine rumen plasmidome. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 5452–5457 (2012).
30. Krawczyk, P. S., Lipinski, L. & Dziembowski, A. PlasFlow: predicting plasmid sequences in

metagenomic data using genome signatures. *Nucleic Acids Res.* **46**, e35 (2018).

31. Pellow, D., Mizrahi, I. & Shamir, R. PlasClass improves plasmid sequence classification. *PLoS Comput. Biol.* **16**, e1007781 (2020).
32. Antipov, D., Raiko, M., Lapidus, A. & Pevzner, P. A. Plasmid detection and assembly in genomic and metagenomic data sets. *Genome Res.* **29**, 961–968 (2019).
33. Hou, S., Cheng, S., Chen, T., Fuhrman, J. A. & Sun, F. DeepMicrobeFinder Sorts Metagenomes into Prokaryotes, Eukaryotes and Viruses, with Marine Applications. *Research Square* (2021) doi:10.21203/rs.3.rs-1016976/v1.
34. Arredondo-Alonso, S., Willems, R. J., van Schaik, W. & Schürch, A. C. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microb Genom* **3**, e000128 (2017).
35. Andreopoulos, W. B. *et al.* Deeplasmid: deep learning accurately separates plasmids from bacterial chromosomes. *Nucleic Acids Res.* **50**, e17 (2022).
36. Zhou, F. & Xu, Y. cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. *Bioinformatics* **26**, 2051–2052 (2010).
37. Carattoli, A. *et al.* In SilicoDetection and Typing of Plasmids using PlasmidFinder and Plasmid Multilocus Sequence Typing. *Antimicrobial Agents and Chemotherapy* vol. 58 3895–3903 (2014).
38. Robertson, J. & Nash, J. H. E. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb Genom* **4**, (2018).
39. Garcillán-Barcia, M. P., Francia, M. V. & de la Cruz, F. The diversity of conjugative relaxases and its application in plasmid classification. *FEMS Microbiol. Rev.* **33**, 657–687 (2009).
40. Rozov, R. *et al.* Recycler: an algorithm for detecting plasmids from de novo assembly graphs. *Bioinformatics* **33**, 475–482 (2017).
41. Pellow, D. *et al.* SCAPP: an algorithm for improved plasmid assembly in metagenomes. *Microbiome* **9**, 144 (2021).
42. Galperin, M. Y., Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.* **43**, D261–9 (2015).
43. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
44. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the

analysis of massive data sets. *Nature Biotechnology* vol. 35 1026–1028 (2017).

45. Fang, Z. *et al.* PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. *Gigascience* **8**, (2019).
46. Schwengers, O. *et al.* Platon: identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores. *Microb Genom* **6**, (2020).
47. Reveillaud, J. *et al.* The Wolbachia mobilome in Culex pipiens includes a putative plasmid. *Nat. Commun.* **10**, 1051 (2019).
48. Wozniak, R. A. F. & Waldor, M. K. Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. *Nat. Rev. Microbiol.* **8**, 552–563 (2010).
49. Wang, G. H. *et al.* Bacteriophage WO Can Mediate Horizontal Gene Transfer in Endosymbiotic Wolbachia Genomes. *Front. Microbiol.* **0**, (2016).
50. Cuecas, A., Kanoksilapatham, W. & Gonzalez, J. M. Evidence of horizontal gene transfer by transposase gene analyses in Fervidobacterium species. *PLoS One* **12**, e0173961 (2017).
51. Garcillán-Barcia, M. P., Alvarado, A. & de la Cruz, F. Identification of bacterial plasmids based on mobility and plasmid population biology. *FEMS Microbiol. Rev.* **35**, 936–956 (2011).
52. Sukupolvi, S. & O'Connor, C. D. TraT lipoprotein, a plasmid-specified mediator of interactions between gram-negative bacteria and their environment. *Microbiol. Rev.* **54**, 331–341 (1990).
53. Norris, S. J., Carter, C. J., Howell, J. K. & Barbour, A. G. Low-passage-associated proteins of Borrelia burgdorferi B31: characterization and molecular cloning of OspD, a surface-exposed, plasmid-encoded lipoprotein. *Infect. Immun.* **60**, 4662–4672 (1992).
54. Jalal, A. S. B. & Le, T. B. K. Bacterial chromosome segregation by the ParABS system. *Open Biol.* **10**, 200097 (2020).
55. Bouet, J.-Y. & Funnell, B. E. Plasmid Localization and Partition in Enterobacteriaceae. *EcoSal Plus* **8**, (2019).
56. Carr, V. R., Shkoporov, A., Hill, C., Mullany, P. & Moyes, D. L. Probing the Mobilome: Discoveries in the Dynamic Microbiome. *Trends Microbiol.* **29**, 158–170 (2021).
57. Meinhardt, F., Schaffrath, R. & Larsen, M. Microbial linear plasmids. *Appl. Microbiol. Biotechnol.* **47**, 329–336 (1997).
58. Chen, Z., Zhong, L., Shen, M., Fang, P. & Qin, Z. Characterization of Streptomyces

plasmid-phage pFP4 and its evolutionary implications. *Plasmid* **68**, 170–178 (2012).

59. Oliva, M. A., Martin-Galiano, A. J., Sakaguchi, Y. & Andreu, J. M. Tubulin homolog TubZ in a phage-encoded partition system. *Proc. Natl. Acad. Sci. U. S. A.* **109**, (2012).
60. Dokland, T. Molecular Piracy: Redirection of Bacteriophage Capsid Assembly by Mobile Genetic Elements. *Viruses* **11**, (2019).
61. Pfeifer, E., Moura de Sousa, J. A., Touchon, M. & Rocha, E. P. C. Bacteria have numerous distinctive groups of phage–plasmids with conserved phage and variable plasmid gene repertoires. *Nucleic Acids Res.* **49**, 2655–2673 (2021).
62. Cury, J., Touchon, M. & Rocha, E. P. C. Integrative and conjugative elements and their hosts: composition, distribution and organization. *Nucleic Acids Res.* **45**, 8943–8956 (2017).
63. Osborn, A. M. & Böltner, D. When phage, plasmids, and transposons collide: genomic islands, and conjugative- and mobilizable-transposons as a mosaic continuum. *Plasmid* **48**, 202–212 (2002).
64. Carraro, N., Poulin, D. & Burrus, V. Replication and Active Partition of Integrative and Conjugative Elements (ICEs) of the SXT/R391 Family: The Line between ICEs and Conjugative Plasmids Is Getting Thinner. *PLoS Genet.* **11**, e1005298 (2015).
65. Vineis, J. H. *et al.* Patient-Specific Bacteroides Genome Variants in Pouchitis. *MBio* **7**, (2016).
66. Monaghan, T. M. *et al.* Metagenomics reveals impact of geography and acute diarrheal disease on the Central Indian human gut microbiome. *Gut Microbes* **12**, 1752605 (2020).
67. Mancabelli, L. *et al.* Meta-analysis of the human gut microbiome from urbanized and pre-agricultural populations. *Environ. Microbiol.* **19**, 1379–1390 (2017).
68. Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649–662.e20 (2019).
69. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).
70. Klümper, U. *et al.* Broad host range plasmids can invade an unexpectedly diverse fraction of a soil bacterial community. *ISME J.* **9**, (2015).
71. Kohler, V., Vaishampayan, A. & Grohmann, E. Broad-host-range Inc18 plasmids: Occurrence, spread and transfer mechanisms. *Plasmid* **99**, 11–21 (2018).
72. Bishé, B., Taton, A. & Golden, J. W. Modification of RSF1010-Based Broad-Host-Range Plasmids for Improved Conjugation and Cyanobacterial Bioprospecting. *iScience* **20**, 216–

228 (2019).

73. Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).
74. Schloss, P. D. Identifying and Overcoming Threats to Reproducibility, Replicability, Robustness, and Generalizability in Microbiome Research. *MBio* **9**, (2018).
75. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]* (2018).
76. Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* **8**, e1002687 (2012).
77. Fernández-López, R. *et al.* Dynamics of the IncW genetic backbone imply general trends in conjugative plasmid evolution. *FEMS Microbiol. Rev.* **30**, 942–966 (2006).
78. Garcillán-Barcia, M. P., Ruiz del Castillo, B., Alvarado, A., de la Cruz, F. & Martínez-Martínez, L. Degenerate primer MOB typing of multiresistant clinical isolates of E. coli uncovers new plasmid backbones. *Plasmid* **77**, 17–27 (2015).
79. Carattoli, A. *et al.* Identification of plasmids by PCR-based replicon typing. *Journal of Microbiological Methods* vol. 63 219–228 (2005).
80. Carattoli, A. *et al.* In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.* **58**, 3895–3903 (2014).
81. Redondo-Salvo, S. *et al.* Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. *Nat. Commun.* **11**, 3602 (2020).
82. Acman, M., van Dorp, L., Santini, J. M. & Balloux, F. Large-scale network analysis captures biological features of bacterial plasmids. *Nat. Commun.* **11**, 1–11 (2020).
83. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
84. Heer, J., Card, S. K. & Landay, J. A. prefuse: a toolkit for interactive information visualization. in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 421–430 (Association for Computing Machinery, 2005).
85. Vrancianu, C. O., Popa, L. I., Bleotu, C. & Chifiriuc, M. C. Targeting Plasmids to Limit Acquisition and Transmission of Antimicrobial Resistance. *Front. Microbiol.* **11**, 761 (2020).
86. MacLean, R. C. & San Millan, A. The evolution of antibiotic resistance. *Science* **365**, 1082–1083 (2019).
87. World Health Organization. *WHO report on surveillance of antibiotic consumption: 2016-2018 early implementation*.

https://apps.who.int/iris/bitstream/handle/10665/277359/9789241514880-eng.pdf (2018).

88. Centers for Disease Control and Prevention. *Antibiotic Use in the United States, 2021 Update: Progress and Opportunities*. https://www.cdc.gov/antibiotic-use/stewardship-report/current.html (2021).
89. Gehrig, S. *et al.* Identification of modifications in microbial, native tRNA that suppress immunostimulatory activity. *J. Exp. Med.* **209**, 225–233 (2012).
90. Galvanin, A. *et al.* Bacterial tRNA 2'-O-methylation is dynamically regulated under stress conditions and modulates innate immune response. *Nucleic Acids Res.* **48**, 12833–12844 (2020).
91. Embers, M. E., Alvarez, X., Ooms, T. & Philipp, M. T. The failure of immune response evasion by linear plasmid 28-1-deficient Borrelia burgdorferi is attributable to persistent expression of an outer surface protein. *Infect. Immun.* **76**, 3984–3991 (2008).
92. Gupta, V. K., Paul, S. & Dutta, C. Geography, Ethnicity or Subsistence-Specific Variations in Human Microbiome Composition and Diversity. *Frontiers in Microbiology* vol. 8 (2017).
93. Obregon-Tito, A. J. *et al.* Subsistence strategies in traditional societies distinguish gut microbiomes. *Nat. Commun.* **6**, 1–9 (2015).
94. Gomez, A. *et al.* Gut Microbiome of Coexisting BaAka Pygmies and Bantu Reflects Gradients of Traditional Subsistence Patterns. *Cell Rep.* **14**, 2142–2153 (2016).
95. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014).
96. Xia, Y., Zhu, Y., Li, Q. & Lu, J. Human gut resistome can be country-specific. *PeerJ* **7**, e6389 (2019).
97. Sonnenburg, E. D. & Sonnenburg, J. L. The ancestral and industrialized gut microbiota and implications for human health. *Nat. Rev. Microbiol.* **17**, 383–390 (2019).
98. Thomas, C. M. Plasmid Incompatibility. *Molecular Life Sciences* 1–3 (2014) doi:10.1007/978-1-4614-6436-5_565-2.
99. Novick, R. P. Plasmid incompatibility. *Microbiological Reviews* vol. 51 381–395 (1987).
100. Velappan, N., Sblattero, D., Chasteen, L., Pavlik, P. & Bradbury, A. R. M. Plasmid incompatibility: more compatible than previously thought? *Protein Eng. Des. Sel.* **20**, 309–313 (2007).
101. Svara, F. & Rankin, D. J. The evolution of plasmid-carried antibiotic resistance. *BMC Evol. Biol.* **11**, 130 (2011).
102. Sykes, R. The 2009 Garrod lecture: the evolution of antimicrobial resistance: a Darwinian perspective. *J. Antimicrob. Chemother.* **65**, 1842–1852 (2010).

103. Cantón, R. & Morosini, M.-I. Emergence and spread of antibiotic resistance following exposure to antibiotics. *FEMS Microbiol. Rev.* **35**, 977–991 (2011).
104. Baquero, F. Low-level antibacterial resistance: a gateway to clinical resistance. *Drug Resist. Updat.* **4**, 93–105 (2001).
105. San Millan, A., Escudero, J. A., Gifford, D. R., Mazel, D. & MacLean, R. C. Multicopy plasmids potentiate the evolution of antibiotic resistance in bacteria. *Nat Ecol Evol* **1**, 10 (2016).
106. Alonso, A., Sanchez, P. & Martinez, J. L. Environmental selection of antibiotic resistance genes. Minireview. *Environmental Microbiology* vol. 3 1–9 (2001).
107. Xiong, W., Sun, Y., Ding, X., Wang, M. & Zeng, Z. Selective pressure of antibiotics on ARGs and bacterial communities in manure-polluted freshwater-sediment microcosms. *Front. Microbiol.* **0**, (2015).
108. Ma, H. & Bryers, J. D. Non-invasive determination of conjugative transfer of plasmids bearing antibiotic-resistance genes in biofilm-bound bacteria: effects of substrate loading and antibiotic selection. *Appl. Microbiol. Biotechnol.* **97**, 317–328 (2012).
109. Berendsen, B. *et al.* Evidence of natural occurrence of the banned antibiotic chloramphenicol in herbs and grass. *Anal. Bioanal. Chem.* **397**, 1955 (2010).
110. Both, L., Botgros, R. & Cavaleri, M. Analysis of licensed over-the-counter (OTC) antibiotics in the European Union and Norway, 2012. *Euro Surveill.* **20**, 30002 (2015).
111. Balbi, H. J. Chloramphenicol: A Review. *Pediatrics in Review* vol. 25 284–288 (2004).
112. Ministry of Health and Medical Services, Government of Fiji. *Fiji Antibiotic Guidelines*. (2019).
113. Delmont, T. O. *et al.* Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nat Microbiol* **3**, 804–813 (2018).
114. van Kessel, M. A. H. J. *et al.* Complete nitrification by a single microorganism. *Nature* **528**, 555–559 (2015).
115. Edwards, R. A. *et al.* Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. *Nat Microbiol* **4**, 1727–1736 (2019).
116. Hug, L. A. *et al.* A new view of the tree of life. *Nat Microbiol* **1**, 16048 (2016).
117. Royer, G. *et al.* PlaScope: a targeted approach to assess the plasmidome from genome assemblies at the species level. *Microb Genom* **4**, (2018).
118. Arredondo-Alonso, S. *et al.* MIplasmids: A user-friendly tool to predict plasmid- and chromosome-derived sequences for single species. *Microb. Genom.* **4**, (2018).
119. Gomi, R., Wyres, K. L. & Holt, K. E. Detection of plasmid contigs in draft genome

assemblies using customized Kraken databases. *Microb Genom* **7**, (2021).
120. Guo, J. *et al.* VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, 37 (2021).
121. Al-Shayeb, B. *et al.* Clades of huge phages from across Earth's ecosystems. *Nature* **578**, 425–431 (2020).
122. Shkoporov, A. N. & Hill, C. Bacteriophages of the Human Gut: The 'Known Unknown' of the Microbiome. *Cell Host Microbe* **25**, 195–209 (2019).
123. Antipov, D., Raiko, M., Lapidus, A. & Pevzner, P. A. Metaviral SPAdes: assembly of viruses from metagenomic data. *Bioinformatics* **36**, 4126–4129 (2020).
124. Smillie, C. S. *et al.* Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480**, 241–244 (2011).
125. Groussin, M. *et al.* Elevated rates of horizontal gene transfer in the industrialized human microbiome. *Cell* **184**, 2053–2067.e18 (2021).
126. Brito, I. L. *et al.* Mobile genes in the human microbiome are structured from global to individual scales. *Nature* **535**, 435 (2016).
127. Galata, V., Fehlmann, T., Backes, C. & Keller, A. PLSDB: a resource of complete bacterial plasmids. *Nucleic Acids Res.* **47**, D195–D202 (2019).
128. Shaiber, A. *et al.* Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. *Genome Biol.* **21**, 292 (2020).
129. Eren, A. M. *et al.* Community-led, integrated, reproducible multi-omics with anvi'o. *Nat Microbiol* **6**, 3–6 (2021).
130. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).
131. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
132. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
133. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
134. Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nat. Commun.* **9**, 1–8 (2018).
135. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
136. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

137. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
138. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12**, 2825–2830 (2011).
139. Feng, Q. *et al.* Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nat. Commun.* **6**, 1–13 (2015).
140. David, L. A. *et al.* Gut Microbial Succession Follows Acute Secretory Diarrhea in Humans. *MBio* **6**, (2015).
141. Raymond, F. *et al.* The initial state of the human gut microbiome determines its reshaping by antibiotics. *ISME J.* **10**, 707 (2016).
142. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
143. Wen, C. *et al.* Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis. *Genome Biol.* **18**, (2017).
144. Le Chatelier, E. *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, (2013).
145. Xie, H. *et al.* Shotgun Metagenomics of 250 Adult Twins Reveals Genetic and Environmental Impacts on the Gut Microbiome. *Cell systems* **3**, 572 (2016).
146. Zeevi, D. *et al.* Personalized Nutrition by Prediction of Glycemic Responses. *Cell* **163**, 1079–1094 (2015).
147. Rampelli, S. *et al.* Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota. *Curr. Biol.* **25**, 1682–1693 (2015).
148. Liu, W. *et al.* Unique Features of Ethnic Mongolian Gut Microbiome revealed by metagenomic analysis. *Sci. Rep.* **6**, (2016).
149. Turnbaugh, P. J. *et al.* The Human Microbiome Project. *Nature* **449**, 804–810 (2007).
150. Shaiber, A. & Murat Eren, A. Anvi'o snakemake workflows. http://merenlab.org/2018/07/09/anvio-snakemake-workflows/ (2018).
151. Murat Eren, A., Vineis, J. H., Morrison, H. G. & Sogin, M. L. A Filtering Method to Generate High Quality Short Reads Using Illumina Paired-End Technology. *PLoS One* **8**, e66643 (2013).
152. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).
153. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*

9, 357 (2012).

154. Jørgensen, T. S., Xu, Z., Hansen, M. A., Sørensen, S. J. & Hansen, L. H. Hundreds of circular novel plasmids and DNA elements identified in a rat cecum metamobilome. *PLoS One* **9**, e87924 (2014).
155. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 1–13 (2019).
156. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* **3**, e104 (2017).
157. Watts, S. C., Ritchie, S. C., Inouye, M. & Holt, K. E. FastSpar: rapid and scalable correlation estimation for compositional data. *Bioinformatics* **35**, 1064–1066 (2018).
158. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).
159. Csardi, G., Nepusz, T. & Others. The igraph software package for complex network research. *InterJournal, complex systems* **1695**, 1–9 (2006).
160. Leiserson, M. D. M., Wu, H.-T., Vandin, F. & Raphael, B. J. CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer. *Genome Biol.* **16**, 160 (2015).
161. Gibson, M. K., Forsberg, K. J. & Dantas, G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.* **9**, 207–216 (2015).
162. Alcock, B. P. *et al.* CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **48**, D517–D525 (2020).
163. Trigodet, F. *et al.* High molecular weight DNA extraction strategies for long-read sequencing of complex metagenomes. *bioRxiv* 2021.03.03.433801 (2021) doi:10.1101/2021.03.03.433801.
164. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19**, 455 (2012).

Data availability

We have also made supplementary data publicly available on zenodo.org. All analyses of reference plasmids and chromosomes are at doi:10.5281/zenodo.5732024. The PlasX model and all analyses of known and predicted plasmids are at doi:10.5281/zenodo.5732447. For all metagenomes, we have compiled the contigs, taxonomic abundances, and PlasX scores at doi:10.5281/zenodo.5730607, gene calls at doi:10.5281/zenodo.5730987, and gene annotations at doi:10.5281/zenodo.5731658. We have deposited long and short sequencing reads from *B. fragilis* isolates into the NCBI Sequence Read Archive (PRJNA782184).

Code availability

We have released two open-source packages, PlasX (https://github.com/michaelkyu/plasx) and MobMess (https://github.com/michaelkyu/mobmess).

Acknowledgments

We thank Karen Lolans (ORCiD:0000-0003-1903-756X) for performing the long-read sequencing and for providing feedback on the manuscript. We thank Samuel Miller (0000-0002-2836-1401) and Marcus Foo (0000-0003-3436-1632) for their insights into tRNA modification genes. We also thank other members of the Meren Lab at the University of Chicago for their feedback. MKY acknowledges support from Toyota Technological Institute at Chicago.

Funding

Center for Data and Computing, at the University of Chicago (MKY, AME)
National Institutes of Health NIDDK grant RC2 DK122394 (AME)
Simons Foundation grant #687269 (AME)
Sloan Foundation (AME)

Author information

Contributions

Conceptualization: MKY, AME

Methodology: MKY, ECF, AME

Investigation: MKY, ECF, AME

Visualization: MKY, ECF, AME

Funding acquisition: MKY, AME

Project administration: MKY, AME

Supervision: MKY, AME

Writing – original draft: MKY, ECF

Writing – review & editing: MKY, ECF, AME

Data curation: MKY, ECF

Formal Analysis: MKY, ECF

Resources: ECF, MKY

Software: MKY

Validation: ECF

Ethics declarations

Competing interests

The authors declare no competing interests.

Supplemental Figures

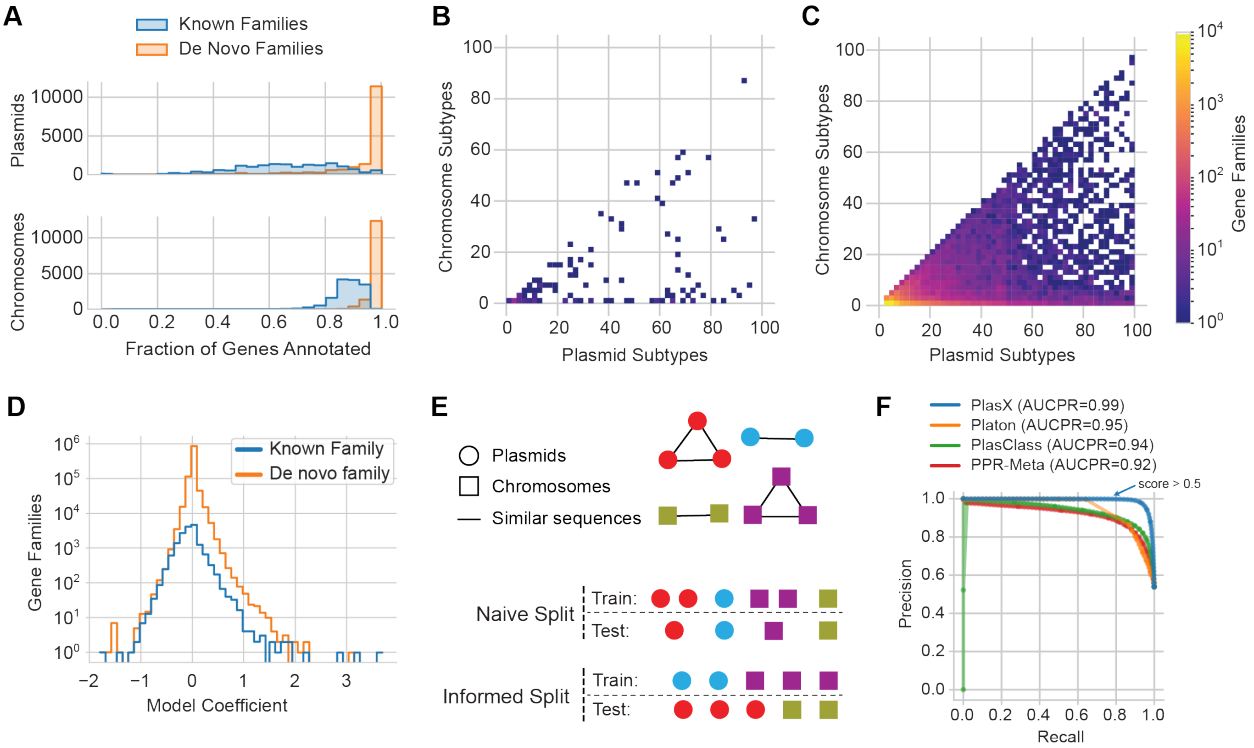


Figure S1. Additional analysis of PlasX. (A) Histograms of reference sequences, based on the fraction of genes that have known or *de novo* family annotations. **(B-C)** Two-dimensional histograms of known (B) and *de novo* (C) gene families, based on the number of plasmid and chromosomal subtypes that each family is found in. The number of gene families is log-scaled. Only the gene families that are enriched in plasmid subtypes (i.e. bottom-right triangle) are shown. **(D)** Histograms of the coefficients learned by PlasX, showing that the vast majority of coefficients are close to zero. **(E)** Diagrams of different training-test split configurations for cross-validation. A random 'naive' split of plasmids and chromosomal sequences results in training and test sets that have similar sequences, due to the existence of plasmid and chromosomal subtypes that contain highly similar sequences. An 'informed' split assigns all sequences of the same subtype to either training or test, creating a more representative evaluation of a model's ability to generalize to unseen sequences. Colors and edges represent sequences that are in the same subtype. **(F)** Precision-recall curves using 4-fold cross-validation and a naive split.

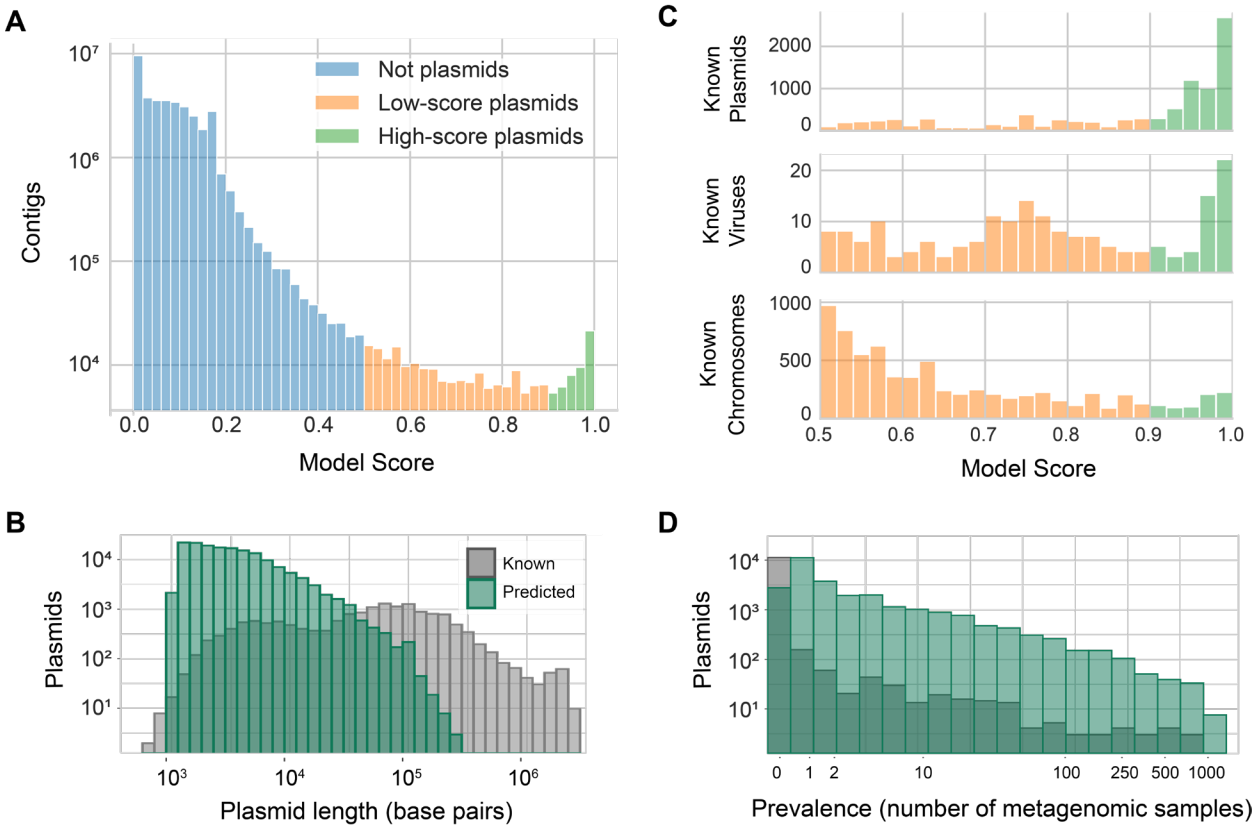


Figure S2. Additional analysis of predicted plasmids. (A) Model scores of all contigs assembled from all 1,782 metagenomes. 226,194 plasmids were predicted by applying a score threshold of >0.5. Of these, 50,163 plasmids were high-scoring (≥0.9 score). **(B)** The sequence length of known and predicted plasmids. **(C)** Model scores of predicted plasmids that matched a sequence in NCBI (≥90% alignment identity and ≥90% coverage of the predicted plasmid). Predictions are labeled as a known 'plasmid', 'virus', or 'chromosome' based on the presence of these words in the description of the matching NCBI sequence. We searched NCBI for only the filtered set of 100,719 non-fragment predictions. **(D)** The prevalence of reference and predicted plasmids across all metagenomes.

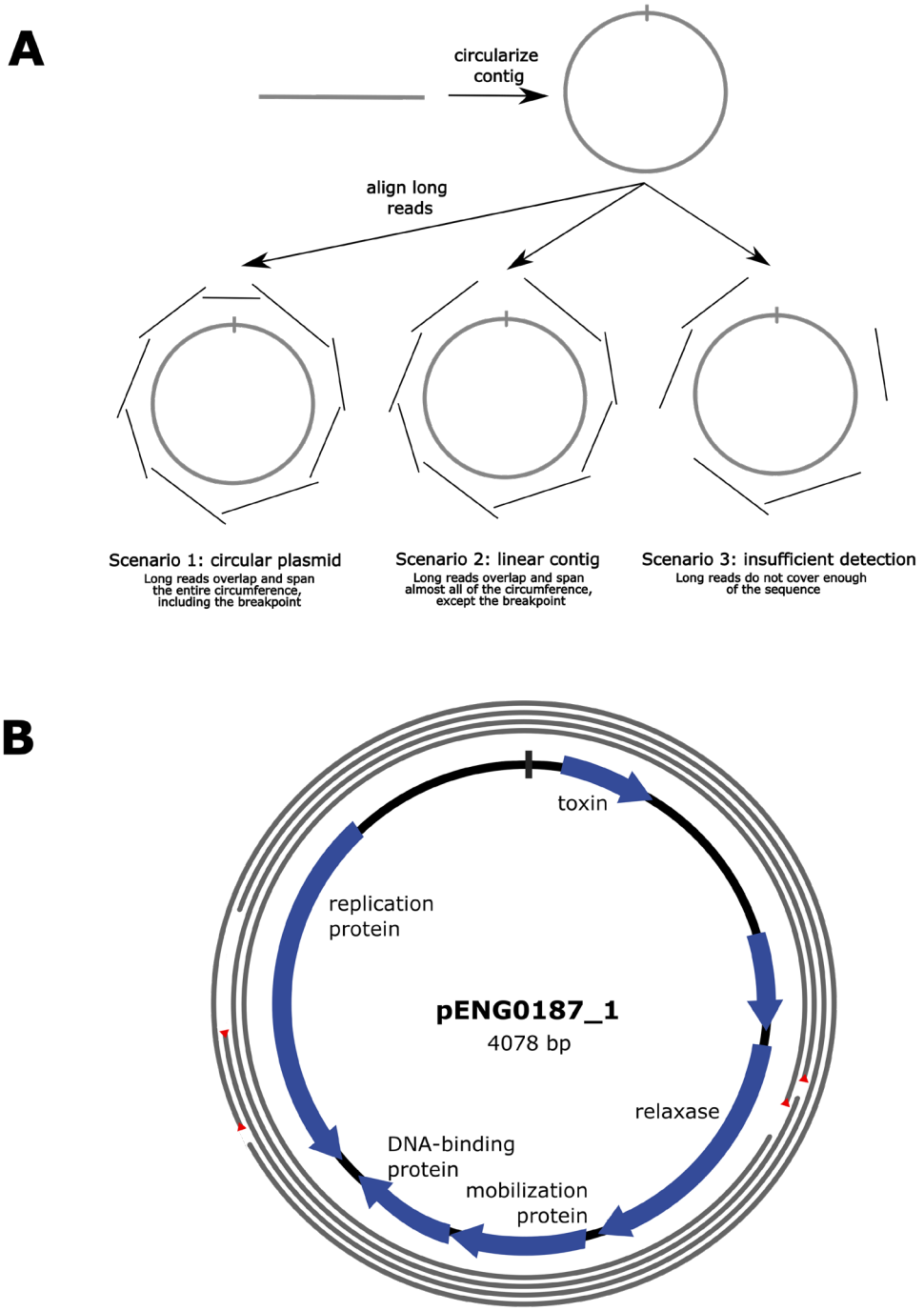


Figure S3. Long read circularity. (A) The process to identify circular plasmids using long read sequences. Contigs are always assembled as linear sequences even when originally circular in the environment. We can determine their original configuration by aligning long reads around the entire sequence. **(B)** *B. fragilis* 216 long reads aligned to pENG0187_1, demonstrating circularity. 4 of 500 reads are shown for simplicity. Red triangles designate the beginning of a long read.

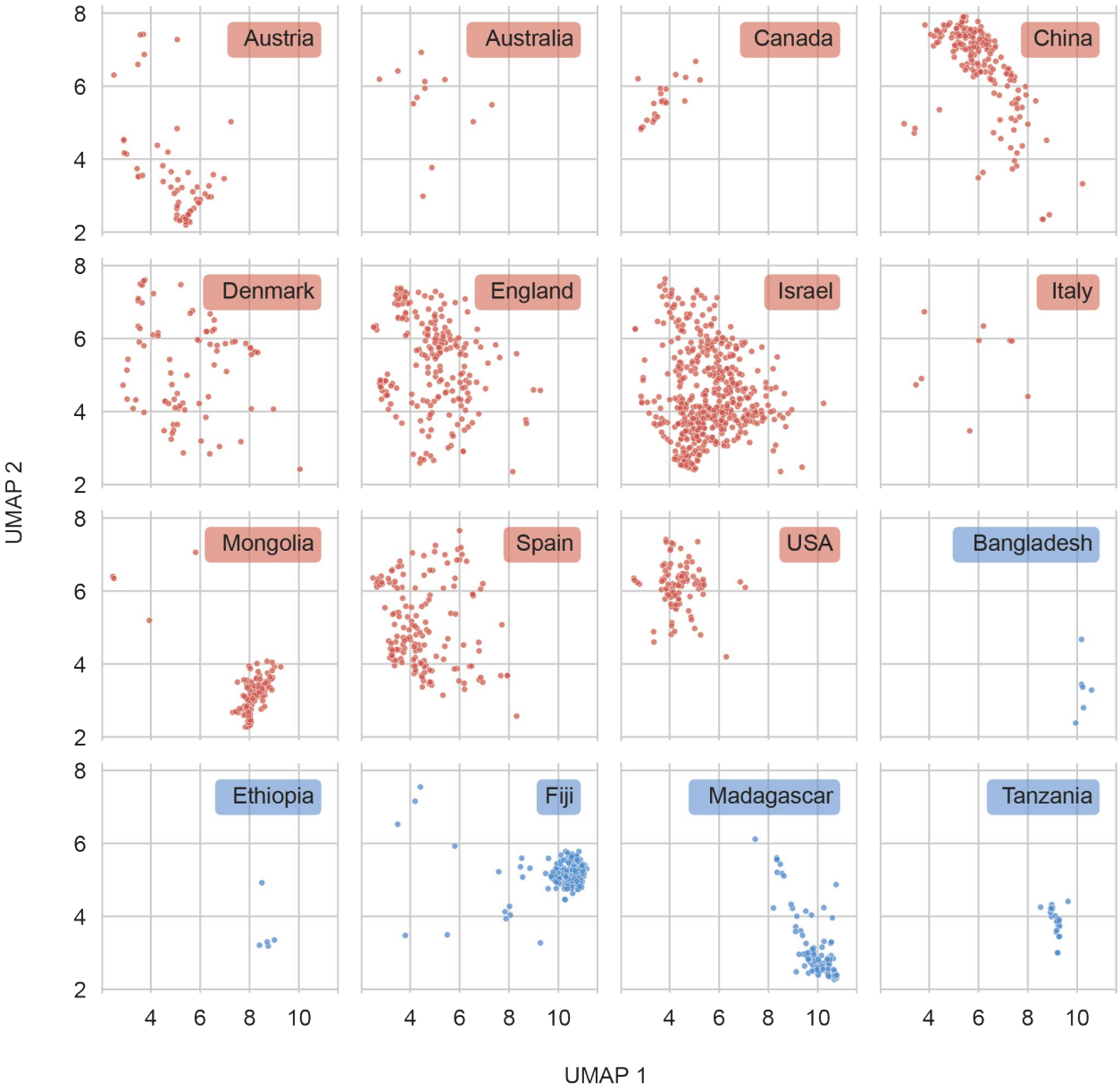


Figure S4. UMAP plot as in Figure 4C. Metagenomes have been partitioned to show clustering within each country.

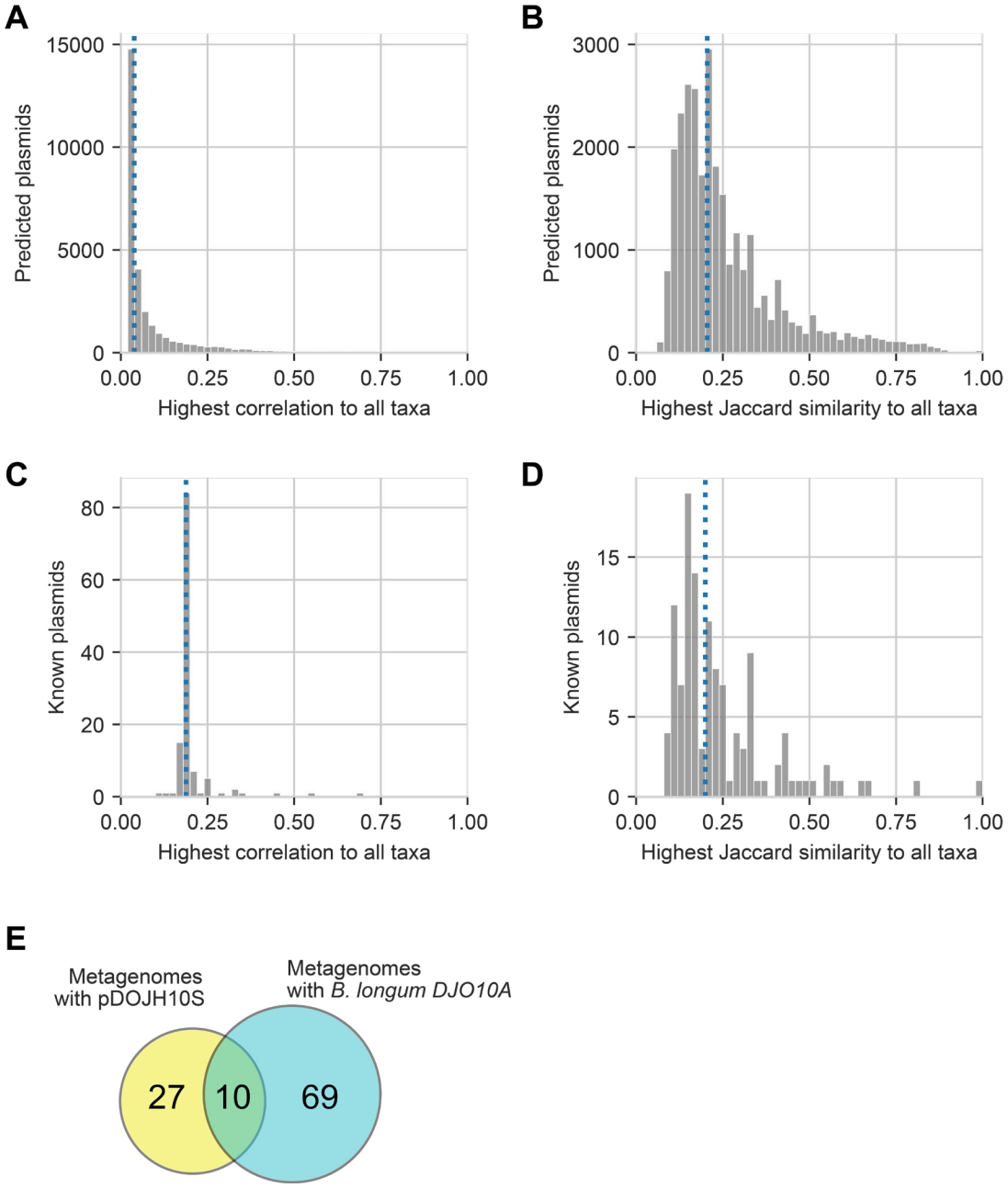


Figure S5. Comparison of the ecological distributions of plasmids and microbial taxonomy. We measured the association between every plasmid and taxon by calculating the correlation between their abundance levels across metagenomes, using the SparCC technique[76]. As another association measure, we applied thresholds to the abundance levels and then calculated the Jaccard similarity between the metagenomes containing the plasmid versus those containing the taxon. We restricted analyses to plasmids that were present in at least 5 metagenomes. **(A-B)** For every predicted plasmid, we identified the taxon with the highest correlation **(A)** or Jaccard similarity **(B)**. **(C-D)** We did the same to identify the best matching taxa of reference plasmids. Blue lines indicate the median of each distribution. **(E)** Venn diagram showing the discordance between the metagenomes containing a plasmid pDOJH10S and those containing its cognate host, a *B. longum* strain.

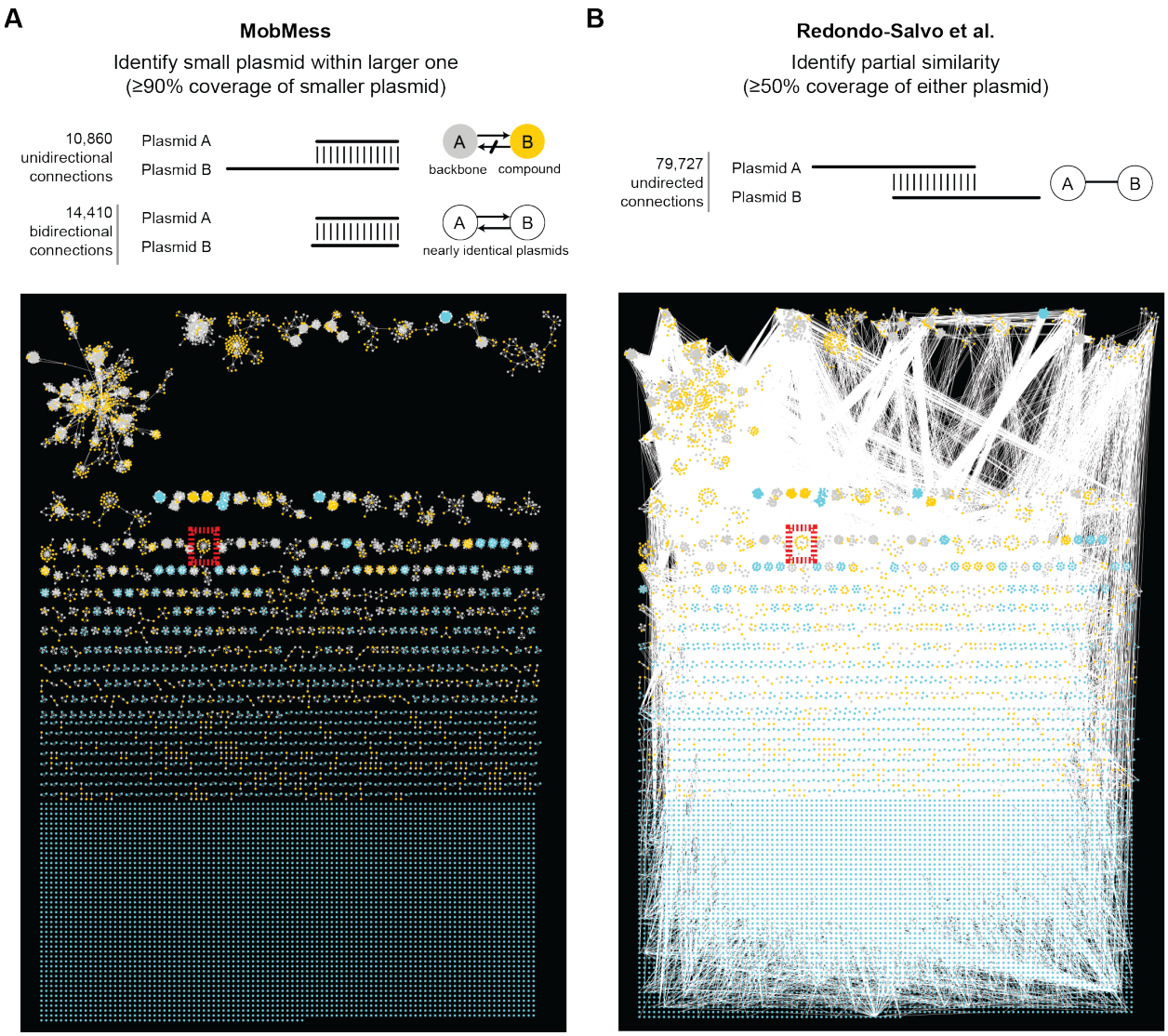


Figure S6. Conceptual differences in constructing plasmid similarity networks. We ran MobMess on the set of 9,894 reference plasmids analyzed by Redondo-Salvo et al.[81]. MobMess constructs a network with directed edges, by aligning plasmids and determining if one plasmid is found as a subsequence within another. Redondo-Salvo et al. constructs a network with undirected edges, by determining whether two plasmids contain partial homology. **(A-B)** Visualization of the similarity networks. We used Cytoscape[83] and the Prefuse directed layout algorithm[84] to lay out the nodes in the MobMess network (A), and then we applied the same layout to the Redondo-Salvo et al. network (B). The red boxes represent the example shown in Figure S7.

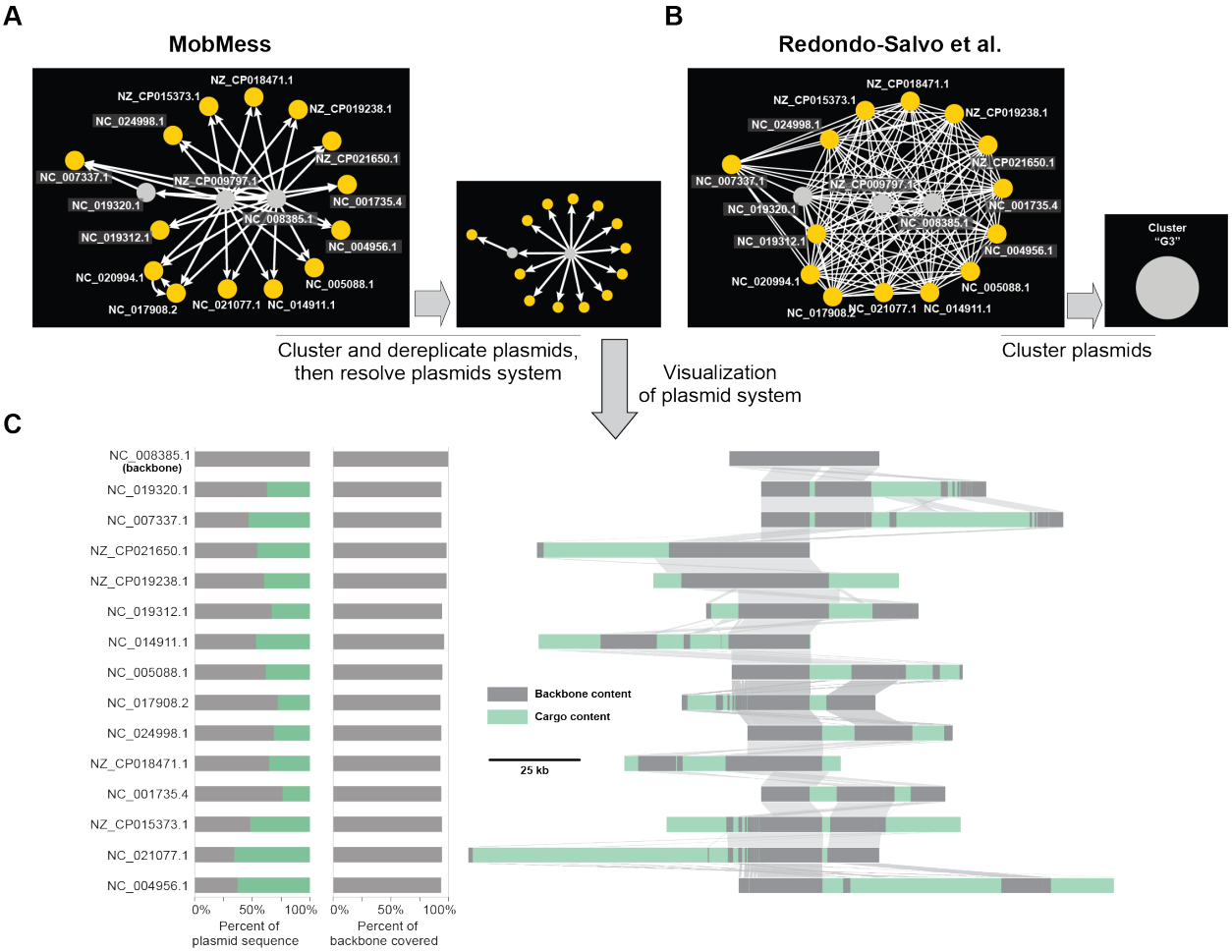


Figure S7. Comparison of MobMess versus Redondo-Salvo et al. for studying a plasmid system. (A-B) An example from the similarity networks in Figure S6, showing the connections between 17 plasmids from the same plasmid system. MobMess further collapses its network to dereplicate plasmids and reveal the plasmid systems's "star"-like topology, where a backbone connects to its compound plasmids. Redondo-Salvo et al. did recognize that these plasmids are related (represented by a cluster called "G3"), but they connected almost every pair of these plasmids in a "hairball" topology, obfuscating the system's internal organization. **(C)** Alignments of plasmids in the MobMess system. Subregions in every sequence are colored gray or green to represent backbone or cargo content, respectively. Ribbons between sequences represent the alignment of subregions. The barcharts show the total breakdown of each plasmid into backbone versus cargo, as well as the fraction of the backbone sequence ('NC_008385.1') that is found within the plasmid.

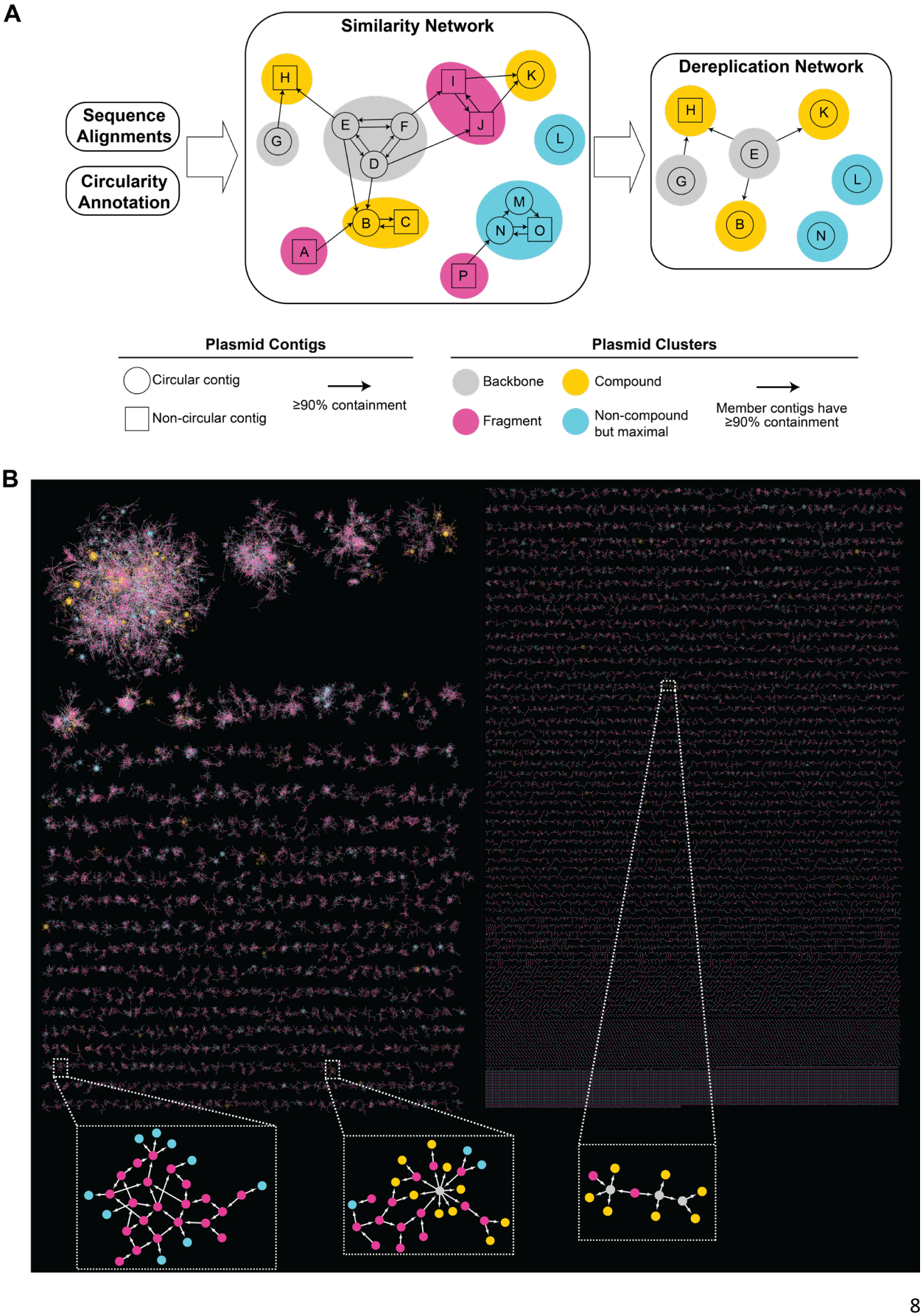

A
Similarity Network
Sequence Alignments
Circularity Annotation
Dereplication Network
H
G
E
F
D
I
J
K
L
B
C
A
M
N
O
P
Plasmid Contigs
Circular contig
Non-circular contig
≥90% containment
Plasmid Clusters
Backbone
Compound
Fragment
Non-compound but maximal
Member contigs have ≥90% containment
B

Figure S8. The MobMess algorithm and application to predicted plasmids. (A) Diagram of the MobMess algorithm for dereplicating plasmids and discovering plasmid systems. All-vs-all sequence alignments and circularity information are used to construct a similarity network of plasmid contigs. Similar contigs are clustered, and every cluster is labeled as either a backbone, fragment, compound, or non-compound maximal. A plasmid system consists of a backbone cluster and the compound clusters connected to the backbone. This example shows two systems: one system has G as the backbone (H is the compound plasmid), and another system has D, E, and F as the backbone (B, C, H, and K are the compound plasmids). To dereplicate, fragment clusters are discarded and a representative sequence is chosen for every non-fragment cluster. **(B)** Network of clusters of predicted plasmids. All clusters are shown except those that are not connected to any other cluster.

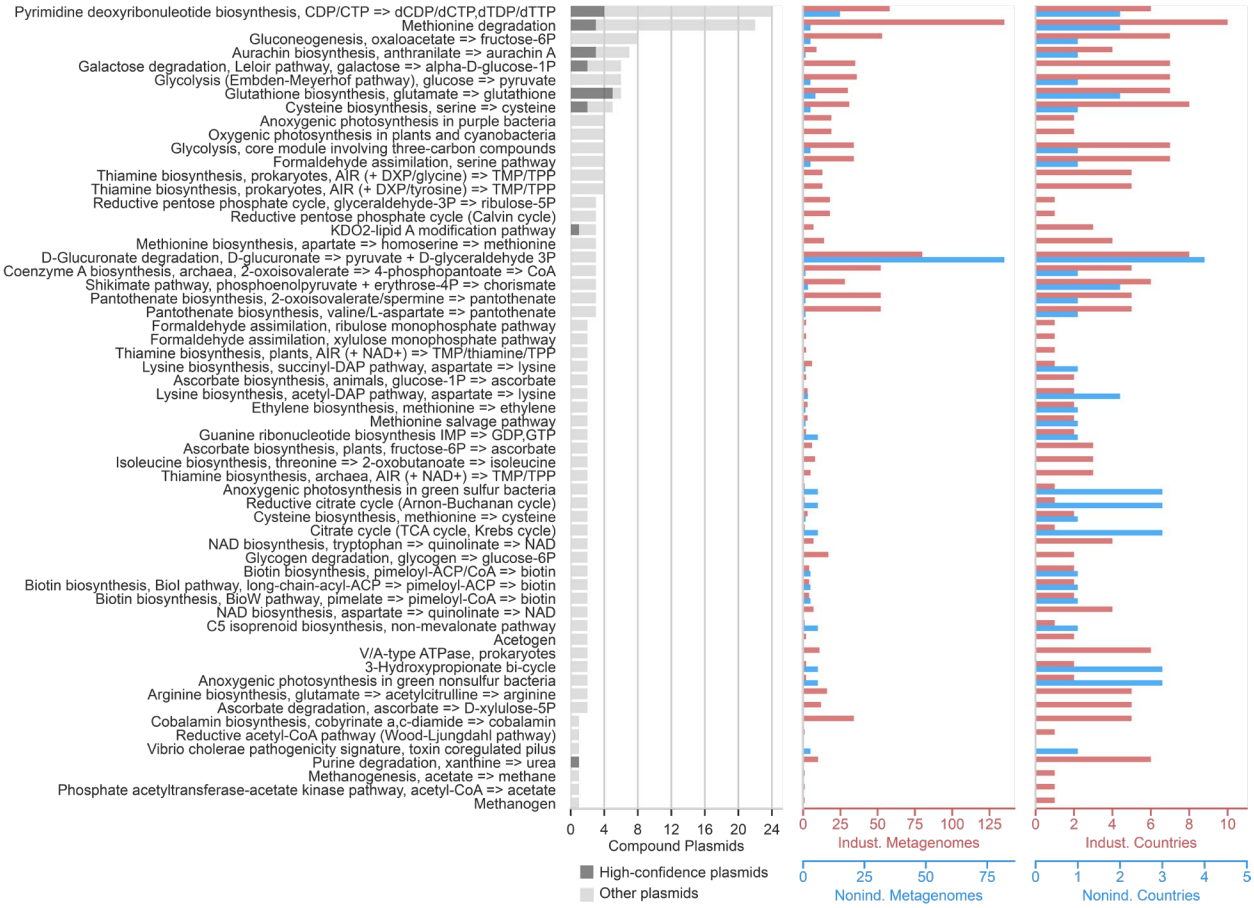


Figure S9. Functional annotation of cargo genes to KEGG modules, similar to Figures 6A and 6B. This plot excludes KEGG modules that occur in only one plasmid system. To avoid redundancy with Figure 6A, this plot also excludes modules that occur in cargo genes annotated to antibiotic resistance.

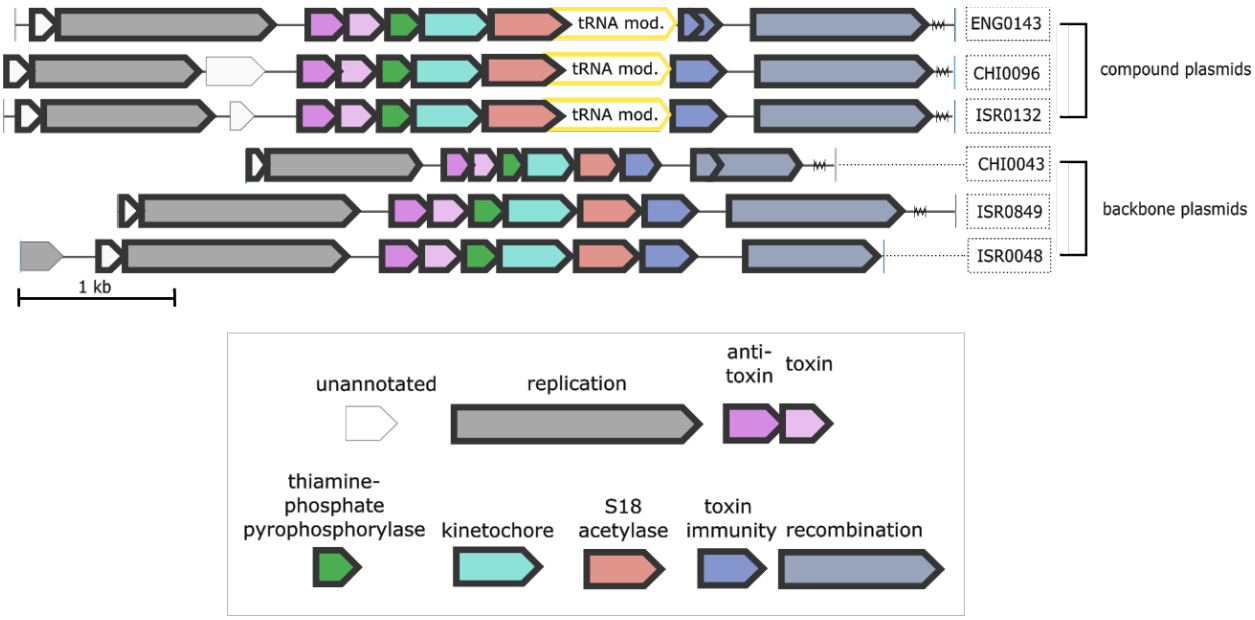


Figure S10. Plasmid system PS1110. Compound plasmids in this system contain a gene that encodes two enzymes, a tRNA Gm18 2'-O-methylase (yellow, 'tRNA mod.') and a Ribosomal protein S18 acetylase (red). Backbone plasmids contain a similar gene that encodes the S18 acetylase but lacks the tRNA methylase. Backbone genes have a thick, black outline.

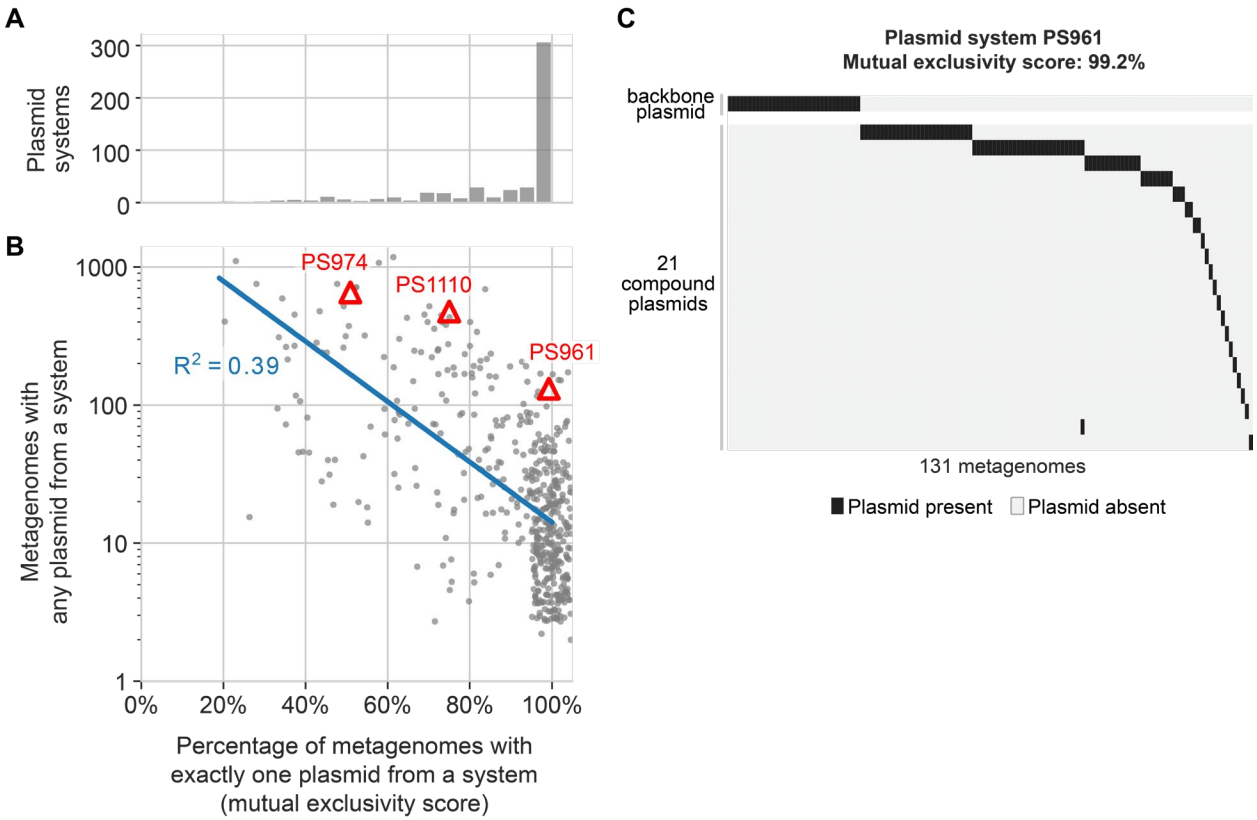


Figure S11. Mutual exclusivity of plasmids in the same system. For every plasmid system with two or more compound plasmids, we defined a 'mutual exclusivity score' to quantify how often its plasmids segregated to different metagenomes. We definethis score as the number of metagenomes that have exactly one of the system's plasmids divided by the number of metagenomes that have any of the system's plasmids. **(A)** Histogram of mutual exclusivity. **(B)** The inverse relation between the mutual exclusivity and prevalence of a system. Red triangles represent examples of plasmid systems that are highly prevalent (present in >100 metagenomes) but are more mutually exclusive than expected by a linear regression (blue line). For easier visualization, the x- and y-coordinates of systems with >90% mutual exclusivity were randomly jittered within ±5% of the axes lengths. **(C)** Visualization of mutual exclusivity in PS961, which is one of the examples in B.

Supplemental Tables

Table S1.

Summary of reference plasmids and chromosomes

Table S2.

Names, accession numbers, and metadata of metagenomes

Table S3.

Summary of predicted plasmids (model scores, orthogonal support, circularity, and NCBI blast results)

Table S4.

Summary of plasmid systems

Table S5.

Gene sequences for plasmid systems shown in Figure 5D, Figure S10 and Figure 6F

Table S6.

DNA extraction and sequencing parameters for long read sequencing of isolate genomes

Table S7.

COGs and Pfams ranked by their PlasX coefficients

Table S8.

Prediction of a Wolbachia plasmid